INTERPRETING DNA EVIDENCE

INTERPRETING DNA EVIDENCE

Statistical Genetics for Forensic Scientists

Ian W. Evett
The Forensic Science Service
United Kingdom

Bruce S. Weir
Department of Statistics
North Carolina State University

To Dennis Lindley and the memory of Clark Cockerham

Contents

In recent months, three different applications have brought home the widespread use and acceptance of DNA evidence. A jury in Santa Monica found a prominent football player responsible for the wrongful death of his former wife and her friend, a New York state crime laboratory was able to identify the remains of all 230 people who perished in an airplane crash, and scientists from the University of California at Davis were able to establish the parentage of the classic grape variety Cabernet Sauvignon. These very different uses of DNA rested on the ubiquity and uniformity of DNA throughout any living organism, and on the near identity of DNA between generations. Of almost equal importance in each case, however, was the numerical reasoning designed to show that matching DNA profiles provided strong evidence in favor of a single source of those profiles. It is this numerical reasoning with which we are concerned in this book. We have assumed some general knowledge of the field of DNA profiling, and would recommend the book *An Introduction to Forensic DNA Analysis* by Inman and Rudin for those without this knowledge.

Writing the book has been both a joy and a challenge. We have enjoyed laying out the foundations of a fascinating field, and we have been gratified by the response to early drafts from participants in the various short courses we have been teaching. We have struggled with the challenge of writing in a way that will be most useful for people who must confront DNA evidence in their professions but who do not have extensive training in statistical genetics. We have each been relatively recently introduced to the other's field, and we have enjoyed the challenge of accommodating each other's different backgrounds and philosophies.

We have benefited by good advice from many colleagues, and we wish to extend special thanks to Dennis Lindley, John Buckleton, Lindsey Foreman, Jim Lambert, Ian Painter, and Charles Brenner. We are very grateful for the comments we received from James Curran, Spencer Muse, Edward Buckler, Dahlia Nielsen, Christopher Basten, Stephanie Monks, Jennifer Shoemaker, and Dmitri Zaykin. Christopher Basten has helped with our LaTeX problems. Andy Sinauer has remained a steady friend to us and our fields. We will welcome comments and suggestions, and we will display any corrections on the World Wide Web at http://www.stat.ncsu.edu (click on "Statistical Genetics"). We can be contacted via that address.

At several places in the text we have illustrated concepts by referring to data collected by the Forensic Science Service in the UK and by Cellmark Diagnostics in the US. We are grateful to both these organizations for their permission to use these data.

Gill Evett and Beth Weir have shown remarkable tolerance for our pre-

occupation with this project over the past few years. This brief mention of our gratitude is all too inadequate.

London and Raleigh
April, 1998

Forensic science is experiencing a period of rapid change because of the dramatic evolution of DNA profiling. The sensitivity and discrimination of techniques now in routine use in many countries were undreamed of ten years ago. DNA has entered the vocabulary of the man in the street, perhaps not so much because of the beautiful work of those such as Watson and Crick as more because of the dramatic impact DNA profiling has had on crime detection.

One of the consequences of this new technology for the forensic scientist is that the strength of the evidence presented at court is usually expressed numerically. Whereas that has been the case for conventional serological techniques for decades, there are now two differences: the scale of implementation of the new methods and the enormous power of the evidence. A match between the profiles of a biological sample from the scene of a crime and that of a suspect has now been shown in many courts in various jurisdictions to have sufficient probative value to convince a jury that the sample came from the suspect, without the need for nonscientific corroborating evidence. The question often asked of a DNA profile is "Is it as good as a fingerprint?" We will meet this question later in the book (like many apparently simple questions, it does not have a simple answer!), but here it gives us an opportunity to reflect on a fascinating paradox.

Many will share the view that DNA profiling is the greatest advance in forensic science since the acceptance of fingerprint identifications by the courts at the turn of the century. Since that time, hundreds of thousands of opinions have been given by fingerprint experts. A fingerprint opinion is never a shade of grey–it is a categorical "those two marks were made by the same finger." This is accepted by courts throughout the world, in most cases without challenge, and the original introduction of this kind of evidence was, apparently, fairly painless. The statistical justification for fingerprint identification was rather sketchy and mainly theoretical.

DNA profiling, on the other hand, received something of a baptism by fire. For a few years, it was conventional to refer to the "DNA controversy," and in some countries there have been long and sometimes bitter courtroom confrontations. The controversy has, in turn, contributed to an explosion in the literature on the subject. Hundreds of papers have been published on DNA profiling statistics, many dealing with data collection, many others dealing with theoretical considerations of probability, statistics, and genetics. Although there were times when the controversy became acrimonious and testifying was unusually stressful, most would now agree that this extended debate has been good for the science. We know that DNA profiling is here to stay and that the statistics of the current techniques have been,

in the main, established as robust. We now can say that we understand far more about the statistics of DNA profiles than about any other forensic technique–including fingerprints!

But another consequence is that there is now an appreciable body of knowledge with which the forensic practitioner must be comfortable if he or she is to report results and give evidence. This book is written with the aim of helping the forensic scientist who works in the DNA profiling field to gain sufficient knowledge of the statistical and genetic issues to report cases and to testify competently. We hope that there will also be much in the book that will interest other groups, such as lawyers and judges, as well as researchers in other fields.

Many forensic scientists engaged in DNA profiling have backgrounds that are strong in the biological sciences but relatively weak in mathematics. We consider that our "core reader" will be a forensic scientist with a degree in one of the biological sciences; thus he or she will be familiar with basic genetics, though we do review the necessary terms and concepts. In our experience, such scientists are often uncomfortable with statistics and so we have deliberately taken a gentle pace over the first three chapters. We have not assumed knowledge of calculus, but we do assume that our readers will be familiar with such mathematical ideas as logarithms, exponentiation, and summation. We must also recognize that there will be other readers with stronger mathematical backgrounds who are interested in a deeper coverage of some of the issues. For the most part, we have worked at keeping the mathematics as simple as possible, and our main aim has been to expose the underlying principles. However, from time to time, we have included more mathematical topics, and wherever possible we have enclosed these in boxes so that the reader who so wishes can skip them.

Because the field is changing rapidly, it has been rather difficult for us to decide what subjects to include and what to leave out. After considerable deliberation we decided to concentrate on the current generation of polymerase chain reaction (PCR) nuclear DNA-based profiling systems. It follows that we have included nothing on the subject of treating measurement error in comparing profiles, so readers will search in vain for a discussion of match/binning. We have not included anything on mitochondrial DNA statistics. That is rapidly becoming a subject in its own right and we have decided, with some reluctance, to defer any treatment of it to a future edition.

We devote all of Chapter 1 to an explanation of basic probability theory because we regard a good grounding in probability to be an essential prerequisite to an understanding of the problems of forensic inference. In

Chapter 2 we show how the Bayesian approach to inference provides a logical and coherent framework for interpreting forensic transfer evidence. In particular, we show how the likelihood ratio is of central importance for forensic interpretation. Chapter 3 covers those topics in basic statistics that are necessary for our purposes; in particular, the theory underlying the estimation of allele proportions, and the basics of classical statistical testing of independence hypotheses. In this chapter we introduce a simple forensic example, based on a rape offense in the imaginary Gotham City, and we use this to discuss both the statistical and population genetics issues involved in assessing the strength of the evidence when the crime profile matches that of a suspect.

Chapter 4 introduces the concepts of population genetics and develops those ideas that are relevant to forensic science. In Chapter 5, we combine the statistical ideas of Chapter 3 with the population genetics from Chapter 4. Chapter 6 is a discussion of inference in cases of disputed parentage. We consider simple paternity and also more complicated situations such as incestuous paternity and identification of human remains. In Chapter 7 we consider the interpretation of cases of profiles of mixtures, illustrated by several examples. In Chapter 8 we return to the Gotham City example introduced in Chapter 3 to illustrate how match probabilities should be calculated, and in Chapter 9 we continue using the example to give our views on how the interpretation of matching profiles should be explained in a statement or formal report. Finally, we talk about the presentation of evidence in court, with particular reference to recent Appeal Court judgments in the UK.

We trust that the reader will appreciate our strategy of using boxes to separate the more mathematical passages. However, we must warn that this approach was inadequate in a few places: indeed, we should have liked to have much of the later sections of Chapters 4 and 5 in boxes!

Chapter 1

Probability Theory

INTRODUCTION

Events, Hypotheses, and Propositions

The word EVENT is often used in the context of probability theory. Basically, we take it to mean any occurrence, past, present, or future, in which we have an interest. Sometimes, in real-life applications when dealing with nonscientists, it appears a little strange. Thus, for example, it is not customary for a lawyer to refer to the event that the defendant assaulted the victim. There are other words we can use in our attempts to understand the world: as scientists we might talk about the truth of a HYPOTHESIS or a PROPOSITION. In the legal context, it is common to use the word ALLEGATION. We will use each of these terms in what follows to mean essentially the same thing; on each occasion we will pick the word that best seems to suit the context.

Randomness

The word RANDOM is used frequently in basic statistics textbooks and also in the forensic context, particularly when the concept of "a person selected at random" is invoked. We will take some time to explain our understanding of the word and the meaning we assign to it when we use it in this book.

When we are talking about DNA types and we talk about "selecting a man at random" we mean that we are going to pick him in such a way as to be as uncertain as possible about his blood type. Another meaning of "random" in this context is selecting a man in such a way that all men have the same chance of being selected. This, however, is a more abstract way of looking at things, particularly when you imagine the practical problems

of actually doing such a selection. Later, we will be talking about human populations in which there is random mating, and in that sense we mean that each person picks a mate in such a way that all of the other members of the population have the same chance of being selected. Clearly this is an abstract idea.

Consider another example. Imagine that we have a chicken's egg before us and we are interested in its weight. Of course, we know something about its weight: it's certainly greater than a milligram and it's certainly less than a kilogram. But as we attempt to make increasingly precise statements, we become increasingly uncertain about their truth. If we can benefit from the experience of having weighed eggs on previous occasions, then we might be able to say that the weight of this particular egg is about 60 g. This represents our best guess, on the information available to us: it might be 50 g, it might be 70 g. The weight of the egg is a random quantity for us.

We can improve our knowledge by actually weighing the egg. Imagine that we use a kitchen scale and it reads 55 g. We now know more but we still can't be certain of the weight. For example, we are not sure about how well the scales have been calibrated, and in any case the scales are graduated only to the nearest 5 g. We now believe that the weight lies between 50 g and 60 g with 55 g our best guess: but the weight itself is still for us a random quantity. The reading on the scales represents data (strictly speaking, a single DATUM) that has enabled us to update our uncertainty about the weight of the egg. The science of statistics is all about using data in this way.

To illustrate our next point, let us consider a simple family board game in which the move of each player in turn is determined by his or her rolling of a single six-sided die. How is that rolling to be done? We could allow each player to hold the die in the hand and place it on the board, but we don't do that because we know that an unscrupulous player, such as five-year-old Suzy, may sneak a look at the die and place it in a way that favors the outcome she needs. So we use a cup, and each player is required to place the die in the cup and tip it out of the cup onto the table. But still the determined unscrupulous player can bend things in his or her favor by placing the die carefully in the cup and sliding it down the side onto the board: the required outcome is not guaranteed but this method appears to favor that which the player seeks, and the other players might feel that it was unfair. The next step then is for the players to agree on a means of manipulating the cup and of ejecting the die from it. What is their requirement? They want the outcome to be fair, and in this sense it means that the outcome should not favor any one player. How do they achieve

this requirement? By making the shaking and tossing processes so chaotic as to make the outcome as unpredictable, or as uncertain, as possible. Note that when we were talking about the egg we were attempting to reduce our uncertainty about something (the weight of the egg), but in this second example we are trying to *maximize* our uncertainty about something (the score that the die will show). If we asked our players what they meant by FAIR, in the present context, they would say that all of the six possible outcomes should have the same probability. Note that we have slipped in this last word before we get to the section on probability and we are not yet talking mathematically. But we don't apologize because we don't think that mathematical concepts are needed here. The players might have used other terms such as "likelihood" or "chance" but their meaning would have been clear to each other and to us. In this context, fairness is equated to the same chance for each outcome.

The process of shaking and tossing is called RANDOMIZATION by statisticians; its sole purpose (other than perhaps to heighten the suspense of the players) is to maximize uncertainty, and this, it turns out, is the best way of achieving fairness. It can be shown using mathematics that maximum uncertainty in the present context is equivalent to the same chance for each outcome; we are not going to give this proof because it is a digression from the main theme of the book, and in our experience most people seem to find our assertion intuitively reasonable.

It is worth saying a little about popular misconceptions relating to randomization, probability, and chance (we are using the latter two words interchangeably at present because in everyday language they are synonymous). There is a tendency to think that the achievement of our belief in equiprobable outcomes in the die throwing example is some physical property of the die. Of course, it is in part because we first establish that the die has a different number on each of its six faces and that a piece of lead has not been craftily inserted immediately behind one face (there wouldn't be much point in using such a die in a board game but it could give an edge in a gambling game such as craps). Given those assurances, our beliefs about the outcome are a consequence of the state of uncertainty we have deliberately created. There is no mysterious physical mechanism–though there is a widespread tendency among both laymen and scientists to speak and act as though there is. Jaynes (1989) calls this the MIND PROJECTION FALLACY. It is a common idea to say that it is the mechanism of chance which ensures that in the long run we end up with equal numbers of each of the six outcomes. But there is no mechanism, no mysterious natural force field governing the behavior of the die. We end up with equal numbers because

we have deliberately maximized our uncertainties. This is the essence of randomization.

We sometimes hear of "the laws of chance" or "the law of averages" as though these are natural laws that govern the behavior of the universe. They are not. They are simply the consequences of our efforts to maximize uncertainty. We will later be talking about the laws of probability, but there we mean something that has a completely different sense from the foregoing popular conception. We will be using probability as a mathematical tool for understanding how we may reason logically in the face of uncertainty. We arrive at the laws of probability in such a way as to make the tools work in such a sensible fashion: they are rather different in nature from, for example, Newton's laws of motion, which are intended to describe aspects of how the natural universe behaves.

We conclude this discussion by referring to an essay by noted mathematician Mark Kac (Kac 1983). Kac pointed out how difficult it is to define randomness and concluded:

> The discussion of randomness belongs to the foundations of statistical methodology and its applicability to empirical sciences. Fortunately, the upper reaches of science are as insensitive to such basic questions as they are to all sorts of other philosophical concerns. Therefore, whatever your views and beliefs on randomness–and they are more likely than not untenable–no great harm will come to you. If the discipline you practice is sufficiently robust, it contains enough checks and balances to keep you from committing errors that might come from the mistaken belief that you *really* know what "random" is.

We believe that the discipline covered in this book is indeed sufficiently robust that the genotypes of samples of people, chosen without knowledge of their genotypes by forensic scientists, can be regarded as random.

PROBABILITY

Uncertainty follows from a deficiency of information and our lives are characterized by decisions that must be made in the face of uncertainty. The issues involved in most real-life decisions are far too complex and fleeting to justify any attempt at logical and numerical analysis. However, there are areas where some progress can be made. Indeed, in the world of science it might be argued that there is a need, wherever possible, to ensure that one's reasoning is governed by logic and quantification rather than emotion and

intuition. This book is concerned with one area of science where there is such a perception.

Probability theory has a history that extends over 200 years, taking its origin in the study of games of chance–the original motivation presumably being that a person who understands decision making in the face of uncertainty will gain an edge over someone who does not. But there have been several different ways of approaching the problem. Each one starts with a different view of probability and how it is defined. Remember that we are talking about a mathematical concept rather than a concrete facet of the natural universe. At the end of the day, the best definition of probability is going to be the one that will take us furthest toward the goal of rational thought, without our losing sight of the real-world problems we are attempting to solve.

Probability Based on Equiprobable Outcomes

Here is an early definition of probability. Think of a hypothetical experiment with several outcomes–the die rolling in the previous section is such an experiment. Now think of an event that is true if one of the outcomes happens. The event "the score is even" is true if the die shows a 2, 4, or 6; the event "the score is greater than 2" is true if the die shows a 3, 4, 5, or 6. Then, if all outcomes are equally probable, the probability of an event H is defined by

$$Probability\ of\ H \quad = \quad \frac{Number\ of\ outcomes\ favorable\ to\ H}{Total\ number\ of\ outcomes}$$

What do we make of this? First of all, it is a definition that can be of considerable use in analyzing many complicated problems, particularly those involving games of chance. It has some serious limitations, however. First, note that it is a definition of probability yet it contains the stipulation that the outcomes must be "equally probable," so the definition is circular. Second, it is restricted in its range of application; indeed, it is useless for most real-life situations. In a criminal trial, for example, a court is concerned with the uncertain event that the defendant committed the crime: there is no possibility of envisioning a number of equally probable outcomes in this situation. We face the same difficulty if we try to invoke the definition to answer a question of the kind "what is the probability that it will rain tomorrow?"

Long-Run Frequency

Defining probability as a long-run frequency is the basis of a school of statistical thought known as FREQUENTIST or CLASSICAL. If we wish to talk about some event H, then it needs to be regarded as the outcome of an experiment that can, in principle at least, be carried out a large number of times. The outcome is assigned a numerical value, or RANDOM VARIABLE, which in this case can take two values, say 1 if H is true and 0 if H is false. We are interested in the number of times that the random variable takes a value equivalent to H being true. Let us carry out N identical experiments (e.g., roll a die N times). If we observe the event H (e.g., the score is an even number) occurring n times then we define the probability of H as the limit of n/N as N approaches infinity. So the probability of H can be determined only by experiment.

This definition can be very useful and, as we have said, most of the modern science of statistics has been built on this foundation. We will be using probabilities assigned on the basis of frequencies a lot later in the book. A useful distinction is to call probability assigned in this way "chance."

However, the frequency definition has limitations. It is intended to make statements about the behavior of random variables; indeed, frequentist probabilities can be applied only to random variables, and the concept of a very long run of random variables is central. This type of probability is then quite different in nature from the probabilities that we talk about in everyday life. If we ask "What is the probability of life on Mars?" then there is no useful purpose in attempting to visualize an infinite collection of planets indistinguishable from Mars. Indeed, this question cannot be answered with a frequentist probability. We face a similar problem when we talk about court cases. The question "What is the probability that this is the defendant's semen?" has only two answers in the frequentist sense: it either is or it isn't, so the probability is either one or zero.

Subjective Probability

A growing number of statisticians have been dissatisfied with the frequency definition of probability and its inferential consequences, and this brings us to the second main school of thought, called BAYESIAN or SUBJECTIVIST. This school recognizes probability simply as a measure of our degree of belief. Although this might sound simple, there are a number of subtle arguments that need to be established before a system of mathematical rigor can be built upon it. We are not going to attempt to get involved in those arguments but refer the interested reader to fundamental works such

as those by O'Hagan (1994) and Bernardo and Smith (1994).

A measurement system should have some kind of calibration standard, and this can be achieved in the following manner as described by Lindley (1991). Let us imagine that we are thinking about whether it will rain tomorrow afternoon at the time we are planning a barbecue. Denote by R the event that rain will spoil tomorrow's barbecue. Imagine a large opaque container that holds 100 balls. They are indistinguishable in size, weight, and shape but there are two colors: black and white. The container is shaken vigorously so that every ball has the same chance of being drawn. We are going to dip in and draw out a ball; we are interested in the probability of the event B that the ball is black. So we now have two uncertain events under consideration: R (rain) and B (black ball). Imagine now that we are going to be given a cash prize if either we correctly predict rain or if we correctly predict a black ball. So we have to choose: which gives us the better chance of winning a prize, predicting R or predicting B? To help us choose the wager we are told the number b of black balls in the container. Now if b is high, say over 90, then the better wager may appear to be that on B (unless we are thinking of a barbecue at a completely unsuitable time of the year). On the other hand, if b is less than 10, then the rain wager may appear preferable. Let us try to think of that value of b at which we are completely indifferent to which wager we are going to choose: assume that we decide that this happens at about $b = 20$. Then our probability of R is $20/100 = 0.2$.

It is this concept of probability that will form the basis of the discussion for the rest of this chapter, although the formal discussion about the laws of probability also holds for frequentist probabilities. Note that as we are talking about subjective (or personal) probabilities we will not talk about them as though they had an independent determinate existence. Instead of talking about determining a probability as we would if we had a frequency definition, we will talk about *assigning* a probability. Given a set of circumstances, you will assign your probability to a particular event, but another person may assign a different probability. Is it a weakness that different people may assign different probabilities to the same event? Not at all–it is an inescapable feature of the real world. We discuss this issue in the next section when we talk about conditioning.

All Probabilities Are Conditional

If you respond "one-half" to the question "What is the probability that this coin will land showing a head?" then think again. Does the coin have two

different sides? How is it to be dropped? Has it been loaded in any way? Only if the answers to these questions are appropriate does it make sense to assign one-half to the probability of a head. But the most important point here is that the probability that we assign depends on what we know: every probability is CONDITIONAL.

In our family board game, if we ask "What is the probability the next throw will be a 3?" then our answer will be conditioned on the following: the die has six sides numbered 1 to 6; the die has not been loaded; the tossing is designed to maximize uncertainty. Only then can we agree on the answer 1/6.

Recall the example of the egg. We weren't actually talking about probability at that time, but we were talking about uncertainty. Our first judgment about the weight of the egg was conditioned by our previous experience about eggs. We could, if we had been asked, have assigned a probability to a proposition of the kind "the weight exceeds 110 g." Later, we learned something new–an item of data–which changed our state of uncertainty and would have undoubtedly changed our probability for the proposition in the previous sentence. The CONDITIONING has changed.

We will be saying much more later about how probabilities change in the light of new data, but at this point we need to digress briefly to talk about three words we will be using frequently: data, information, and evidence. Some writers would not distinguish between these three in a discussion of probability, arguing that probabilities can always be viewed as being conditioned by data. But we believe that understanding can be eased by making distinctions in the following way. We will use DATA when we are referring to observations that can be quantified in some way: for example, they might be a set of genotypes from samples examined by a scientist. We will use INFORMATION to refer to things that are not so easily quantifiable, such as an eyewitness report that the offender was Caucasian. This word has a more general meaning, and if we have a collection of information that includes some data, such as a report that there were three offenders, then we will use the more general term. The word EVIDENCE will be used in a still more general sense to include both data and information, particularly when we are talking about preparing a statement or presenting results in court. The distinctions between the three are not hard and fast and the reader should not feel confused if we sometimes use one of the three rather than another that appears more suitable in a given situation.

In our experience, when scientists disagree about the probability of a proposition, then it is often because they have different conditioning. It may be a simple matter of knowledge; for example, in speculating about the

probability of rain tomorrow, one person may have better local knowledge of conditions than another. In scientific arguments, one proponent might have a slightly different model for reality than another does. It may be a matter of information: in the weather example, one may have heard a different weather forecast from the other. This is the reason that it is necessary always to be as clear as possible about the conditioning.

Notation

We have now reached the stage at which we need to introduce some notation. Choosing notation is always a compromise, but without it arguments become impossibly verbose and even harder to understand. Comprehension can be aided by a carefully chosen notation.

This is our first piece of notation: $\Pr(H|E)$, that is shorthand for "the probability of H given E." Here, H is some event or proposition about which we are uncertain, and E is the evidence (information and/or data) that we have in relation to H. If H is an event, then, depending on context, we might read $\Pr(H|E)$ as

- the probability that H has occurred,
- the probability that H will occur, or
- the probability that H is true.

We will also use $\bar{H}$ to denote the COMPLEMENT of H. Then $\Pr(\bar{H}|E)$ denotes

- the probability that H has not occurred,
- the probability that H will not occur, or
- the probability that H is not true.

Sometimes we will summarize all of the conditioning by one letter, but sometimes it will be necessary to list different aspects. So we might write something like $\Pr(H_p|G_S, G_C, I)$, where

- H_p is the (prosecution) hypothesis that the semen came from the suspect.
- G_S and G_C denote DNA profiles, or genotypes, of the suspect and crime sample, respectively.

- I denotes all of the other evidence relevant to the allegation (what the complainant said, for example).

The commas have no function other than to aid clarity by separating the different items of conditioning. We will use $\Pr(H)$ to denote "the probability of H" in situations in which we are confident that there is no disagreement between all participants in a discussion about the conditioning. This will normally be done when we are talking about abstract problems where the conditioning can be explicitly established beforehand. The abbreviation of $\Pr(H|E)$ to $\Pr(H)$ is again done to assist in achieving clarity.

Now that we have decided on what we mean by probability and settled on some basic notation, it is necessary to state the logical rules that govern the ways in which probabilities may be combined. The requirements of those rules are that they should be as simple as possible, but the results that derive from them must be consistent with rational thought. This is called COHERENCE by some writers (see, for example, Lindley 1982). If a gambler, for example, were using an incoherent set of rules for deciding on his betting strategy, then he would tend to lose money, even if the betting system was fair for him. It turns out that the requirements can be satisfied by three simple rules, or laws. There are various ways of expressing the rules, but we have selected the following way because we hope that it is the easiest to follow.

THE LAWS OF PROBABILITY

The First Law of Probability

The first law is undoubtedly the simplest. It tells us that probability can take values in the range zero to one only and that an event that is certain to occur has probability one. The FIRST LAW OF PROBABILITY is

$$\begin{cases} 0 \leq \Pr(H|E) \leq 1 \\ \Pr(H|H) = 1 \text{ for any } H \end{cases}$$

that is, if we know H is true, then it has a probability of 1.

From this we can later deduce, using the second law, that if a proposition is false, it has a probability zero. Note that it is a common convention to multiply probabilities by 100 and call them PERCENTAGE PROBABILITIES. We will do this ourselves from time to time.

Box 1.1: General form of second law of probability
If $H_i, i = 1, 2, \cdots, r$, are mutually exclusive events given E, then

$$\begin{aligned} \Pr(H_1 \textit{ or } H_2 \textit{ or } \ldots \textit{ or } H_r|E) &= \Pr(H_1|E) + \Pr(H_2|E) + \\ &\quad \cdots + \Pr(H_r|E) \\ &= \sum_i \Pr(H_i|E) \end{aligned}$$

The Second Law of Probability

Return to the example of throwing a die that we talked about earlier. Given all the conditions to which we can agree for a fair tossing, then we also agree that the probability of any one score is 1/6. If we ask the probability of the event that the die will show an even number, few people will hesitate before replying 1/2, and this is the result of adding together the probabilities for a 2, a 4, and a 6. There is one very important feature of these last three events that we must emphasize: they are MUTUALLY EXCLUSIVE. If any one of them occurs, then none of the others has occurred. It is obviously the case with die throwing that each throw can result in only one number. Another feature of the event we stipulated–an even number–was the implication of the word *or*. The event is satisfied by EITHER the event 2, OR the event 4, OR the event 6. This suggests a rule for adding probabilities that is intuitively reasonable: If two events are mutually exclusive and if we wish to know the probability that one or other of them is true then we simply add their probabilities. Formally, if G and H are mutually exclusive events, given E, then the SECOND LAW OF PROBABILITY is

$$\Pr(G \textit{ or } H|E) = \Pr(G|E) + \Pr(H|E)$$

This law is generalized to an arbitrary number of mutually exclusive events in Box 1.1.

There is a useful corollary to the first and second laws. If $\Pr(H|E)$ is the probability that H is true (or that event H occurs) then $\Pr(\bar{H}|E)$ denotes the probability that H is false (or that event H does not occur). Because these two events are mutually exclusive

$$\Pr(H \textit{ or } \bar{H}|E) = \Pr(H|E) + \Pr(\bar{H}|E)$$

they are also EXHAUSTIVE in that between them they cover all possibilities–one or other of them must be true. So,

$$\Pr(H|E) + \Pr(\bar{H}|E) = 1$$

and

$$\Pr(\bar{H}|E) \quad = \quad 1 - \Pr(H|E)$$

The probability that H is false is one minus the probability that H is true. For an event ($\bar{H}$) that is false (when H is true)

$$\Pr(\bar{H}|H) \quad = \quad 1 - \Pr(H|H) = 0$$

as mentioned in the previous section.

Exercise 1.1 When a die is tossed, suppose E is the information that the face showing is even, H_1 is the event that the face shows a 2, and H_2 the event it shows a 4. What are $\Pr(H_1|E)$, $\Pr(H_1 \textit{ or } H_2|E)$, $\Pr(\bar{H}_1|E)$, and $\Pr(H_1|\bar{E})$?

The Third Law of Probability

We lead up to this law by means of two examples. First, suppose there is an opaque jar containing a large number of balls that are indistinguishable in shape, size, and weight: half are black and half are white. We have a balanced coin with a head and a tail. We are going to draw a ball and then toss the coin, and we are interested in the probability that the ball will be black (B) and the coin will land showing a head (H). It seems reasonable to assign the probability 1/2 to B. If we do get a black ball, then what does that tell us about the outcome of the coin tossing? Absolutely nothing, of course. So the probability of H is 1/2 no matter what happens when we draw a ball. It is unlikely that people will disagree that the probability of B and H is the product of these two probabilities: 1/4.

In the next example we have the same jar containing equal numbers of black balls and white balls, but there is no coin tossing. Instead, each ball is marked with a letter: half are marked H and half are marked T. Now we are going to draw a ball, and again we are interested in the probability that B and H will occur. Again we agree to assign the probability 1/2 to B and 1/2 to the probability H. But can we multiply these two numbers together? Well, we could, but we'd be better off asking for information on how the balls have been marked: are all the black balls marked H and all the white balls T? If so, the information that B is true tells us that H must be true also. In the special case where half the black balls are marked H, then the probability of B and H is 1/4, otherwise it is something else. Once we know that B has occurred, we need to know the probability of H given that B has occurred. This gives us two probabilities that we can multiply together.

Box 1.2: General form of third law of probability
For all events $H_i, i = 1, 2, \ldots, r$,

$$\begin{aligned}\Pr(H_1 \textit{ and } H_2 \textit{ and } \ldots \textit{ and } H_r|E) &= \Pr(H_1|H_2, H_3, \ldots, H_r, E) \\ &\quad \times \Pr(H_2|H_3, \ldots, H_r, E) \times \cdots \\ &\quad \times \Pr(H_r|E)\end{aligned}$$

For any two events, J and K, the THIRD LAW OF PROBABILITY can be written:

$$\Pr(J \textit{ and } K|E) = \Pr(J|E)\Pr(K|J, E)$$

A general statement of the law is given in Box 1.2.

For a little while, we will assume that the conditioning information is clearly specified and the same for all of us, and then we present this equation in a slightly less forbidding form:

$$\Pr(J \textit{ and } K) = \Pr(J)\Pr(K|J)$$

There is no particular reason J should precede K, and the law can also be written

$$\Pr(J \textit{ and } K) = \Pr(K)\Pr(J|K)$$

We can illustrate the application of the third law with the following example. Three-quarters of the population of a hypothetical country are Chinese and one-quarter is Asian Indian. We are interested in the following question: if a person is selected from the population, then what is the probability that person will be Asian Indian (event J) with a *HUMTHO1* genotype 8,9.3 (event K)? We agree that the sensible value to assign to $\Pr(J)$ is 0.25; how do we complete the calculation? Clearly, the answer involves the proportion of people who are 8,9.3. But we are not interested in the proportion of the *entire* population with this genotype, or the proportion of the Chinese population who are 8,9.3. Assume that we are told, by someone who has done some research on the locus, that 4.8% of Asian Indians are type 8,9.3; then it is reasonable to assign $\Pr(K|J) = 0.048$. So

$$\begin{aligned}\Pr(J \textit{ and } K) &= \Pr(J)\Pr(K|J) \\ &= 0.25 \times 0.048 = 0.012\end{aligned}$$

We return to this example later.

Exercise 1.2 In another hypothetical country, 80% of the registered voters are Caucasian. Of the Caucasian voters, 20% inhabit the highlands and the remainder the lowlands. Among the Caucasian highlanders, 75% speak an ancient Celtic language. If we select a person at random from the voter registration list, what is the probability that he or she will be a Celtic-speaking, Caucasian highlander?

Independence

There is a special case in which the information that K is true does nothing to change our uncertainty about J (and vice versa). The earlier example with the balls in the container and the coin was such a case. Then $\Pr(J|K) = P(J)$ and, from the first of these two equations,

$$\Pr(J \textit{ and } K) \quad = \quad \Pr(J)\Pr(K)$$

In this special case the two events are said to be statistically INDEPENDENT or UNASSOCIATED.

Note that, although we have omitted E for brevity from our notation, the independence or otherwise of J and K will be determined by E. If every white ball in the urn of the previous example was marked with a T and every black ball was marked with an H, then B and T are dependent. If half of each color are marked H and half marked T, then B and T are independent. It is more correct to say that events are INDEPENDENT CONDITIONAL ON E.

When we discuss forensic transfer evidence later, we will encounter situations where two events are independent under one hypothesis, but dependent under another. Here is an example: in an inheritance dispute a man claims to be the brother of a deceased person. Under his hypothesis, the events that he and the deceased person have a particular DNA profile are dependent, but under the hypothesis that he is unrelated to the deceased person the two events may be taken as independent. As we will see in Chapter 4, sibs are more likely than unrelated people to have the same DNA profile.

The notion of independence extends in a similar way to three or more events, as we shall see later when we combine genotype probabilities across multiple loci.

Box 1.3: Derivation of the law of total probability
Let $S_1, S_2, \ldots, S_r$ be r mutually exclusive events. Furthermore, let them be exhaustive so that $\sum_i \Pr(S_i) = 1$. Let R be any other event. Then the events $(R \textit{ and } S_1), (R \textit{ and } S_2), \ldots, (R \textit{ and } S_r)$ are also mutually exclusive. The event

$$(R \textit{ and } S_1) \textit{ or } (R \textit{ and } S_2) \textit{ or } \cdots \textit{ or } (R \textit{ and } S_r)$$

is simply R, because the S_i are exhaustive. So, from the second law

$$\Pr(R) \quad = \quad \Pr(R \textit{ and } S_1) + \Pr(R \textit{ and } S_2) + \ldots + \Pr(R \textit{ and } S_r)$$

Then, by the third law

$$\Pr(R) \quad = \quad \sum_i \Pr(R|S_i) \Pr(S_i)$$

The Law of Total Probability

From the above three laws follow the entire theory of probability; no further basic laws are needed. However, there are certain standard results that are used frequently. The first of these, the law of total probability, has also been called the "law of the extension of the conversation" (see, for example, Lindley 1991). If A and B are two mutually exclusive and exhaustive events (so that $B = \bar{A}$), then for any other event H, the LAW OF TOTAL PROBABILITY states that

$$\Pr(H) \quad = \quad \Pr(H|A) \Pr(A) + \Pr(H|B) \Pr(B)$$

It is derived in a more general form from the first three laws as shown in Box 1.3, and readers who wish to do so may skip the derivation.

We use this result when we are interested in evaluating the probability of an event that depends on a number of other events that are themselves mutually exclusive. The examples we do in Chapter 6 on family trees and parentage analysis illustrate this, but a small foretaste may help. Let us imagine that we are interested in determining the probability that one of the *HUMTHO1* alleles an individual inherits is allele 8. One way of looking at this is to say that *either* the mother had the 8 allele and passed it on to the offspring *or* the father had the 8 allele and passed it on. We won't work that calculation at this stage, because it is complicated a little by the need to take account of the possibilities of homozygosity in the parents. Instead we give an example that is based on one by Berry (1996). Fred is playing in a chess tournament. He has just won his match in the first round. In

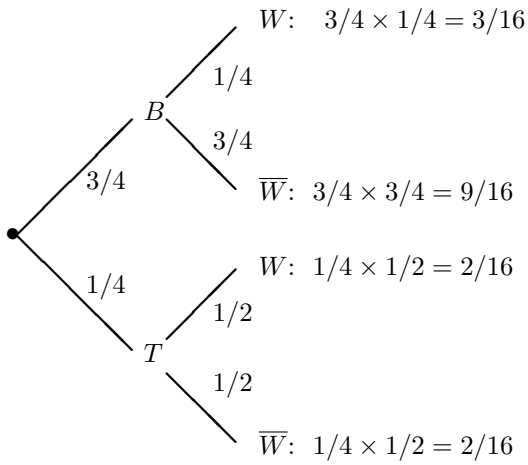


Figure 1.1: Tree diagram for chess tournament.

the second round he is due to play either Bernard or Tessa depending on which of those two wins their first round match. Knowing what he does about their respective abilities, he assigns 3/4 to the probability that he will have to play Bernard, $\Pr(B)$, and 1/4 to the probability that he will have to play Tessa, $\Pr(T)$. We should write these as $\Pr(B|K)$ and $\Pr(T|K)$, where K denotes Fred's knowledge of their respective abilities, but we leave the conditioning out just to simplify the notation.

What is the probability of W, the event that Fred will win? He assigns 1/4 to the probability that he will beat Bernard if he has to play him, $\Pr(W|B)$, and 1/2 to the probability that he will beat Tessa if he has to play her, $\Pr(W|T)$. So, the probability that he will win his second round match, $\Pr(W)$, is the probability of Bernard winning and Fred beating Bernard plus the probability of Tessa winning and Fred beating Tessa. That is:

$$\begin{aligned}\Pr(W) &= \Pr(W|B)\Pr(B) + \Pr(W|T)\Pr(T) \\ &= \left(\frac{1}{4}\times\frac{3}{4}\right) + \left(\frac{1}{2}\times\frac{1}{4}\right) \\ &= \frac{5}{16}\end{aligned}$$

It is useful to present this type of analysis as the tree diagram shown in Figure 1.1.

Exercise 1.3 According to the 1991 census, the New Zealand population con-

sists of 83.47% Caucasians, 12.19% Maoris, and 4.34% Pacific Islanders. The probabilities of finding the same *YNH24* genotype as in the crime sample in the case *R. v. Ladbrook* from a Caucasian, Maori, or Pacific Islander are 0.013, 0.045, and 0.039 respectively. What is the probability of finding that genotype in a person taken at random from the whole population of New Zealand?

Odds

Before we explain another important result of probability theory, we need to explain the notion of ODDS. It is a term used in betting, where it means something slightly different from what it means in formal theory. In everyday speech, odds and probability tend to be used interchangeably; this is a bad practice because they are not the same thing at all.

If we have some event H about which the conditioning is unambiguous, then $\Pr(H)$ denotes the probability of H, and the odds in favor of H are

$$O(H) \quad = \quad \frac{\Pr(H)}{\Pr(\bar{H})}$$

i.e.,

$$O(H) \quad = \quad \frac{\Pr(H)}{1 - \Pr(H)}$$

Because $O(H)$ is the ratio of two probabilities, it can take any value between zero (when H is false) and infinity (when H is true). Let us consider some numerical examples. In the family board game, given the agreed conditions for rolling the die, we have 1/6 as our probability for the event that the die will show a score 3, and 5/6 for the probability that it will not show a score 3. The odds in favor of a 3 showing are:

$$O(3) \quad = \quad \frac{1/6}{5/6} = \frac{1}{5}$$

When, as here, the odds are less than one, it is customary to invert them and call them ODDS AGAINST. So, in this case, the odds are 5 to 1 against.

Now consider a fair coin toss example where we can agree that the probability of a head $\Pr(H)$ is 0.5. Then the odds in favor of a head are 0.5/0.5 = 1. Odds of one are conventionally called EVENS, no doubt as an indication that things are evenly balanced.

For a last example on converting probabilities into odds, think of the first round of the chess game described earlier. Fred's probability that Bernard will beat Tessa, $\Pr(B)$, is 3/4 so the odds in favor of Bernard winning are

3/4 divided by 1/4 or 3. When odds are bigger than one it is conventional to call them ODDS ON so in this case the odds are 3 to 1 on that Bernard will beat Tessa (that is, as far as Fred is concerned–Bernard and Tessa may each have a different view!).

It is our experience that people have little difficulty in converting from probabilities to odds, but converting from odds to probabilities sometimes causes problems. The simplest solution is to remember a formula that can easily be derived by rearranging the above expression for $O(H)$:

$$\Pr(H) \quad = \quad \frac{O(H)}{1+O(H)}$$

Exercise 1.4 Two regular six-sided dice are rolled fairly. (a) What are the odds that they both show an even number? (b) What are the odds that they both show a six?

Exercise 1.5 Convert the following odds in favor of H to the probabilities of H: (a) $O(H) = 19; (b)$O(H)=0.2; (c)O(H)=1000; (d) $O(H) = 1/1000$.

BAYES' THEOREM

Bayes' Theorem is a very important result that we will be using frequently. It is attributed to an 18th century clergyman, Thomas Bayes, and is now recognized as a useful model for understanding how evidence may be presented logically and impartially in legal proceedings.

In the problems we will be discussing in later chapters there will generally be some event or proposition H about which there is uncertainty and some conditioning information I. Then there is some additional information E that, in our problems, will generally be the DNA evidence, and we are interested in how this affects the uncertainty about H. Bayes' theorem provides a model for doing this. If H_p and H_d are the prosecution and defense hypotheses, one form of the theorem is as follows:

$$\frac{\Pr(H_p|E,I)}{\Pr(H_d|E,I)} \quad = \quad \frac{\Pr(E|H_p,I)}{\Pr(E|H_d,I)} \times \frac{\Pr(H_p|I)}{\Pr(H_d|I)}$$

This is the ODDS FORM OF BAYES' THEOREM and its proof is contained in Box 1.4. We will illustrate the way it can be used by means of an example.

Let us return to the example where there were black and white balls in an opaque container. This time we are told that one-third are black and two-thirds are white. As before, some of the balls carry the letter H and some the letter T, but now we are given the following information:

Box 1.4: Derivation of Bayes' theorem
Let H_p be an event or proposition, and let E be an item of evidence. Let I denote all the relevant background information. Then the third law gives

$$\begin{aligned} \Pr(H_p \textit{ and } E|I) &= \Pr(H_p|I)\Pr(E|H_p, I) \\ \Pr(H_p \textit{ and } E|I) &= \Pr(E|I)\Pr(H_p|E, I) \end{aligned}$$

Equating the right hand sides of these two expressions

$$\Pr(H_p|I)\Pr(E|H_p, I) = \Pr(E|I)\Pr(H_p|E, I)$$

and, rearranging,

$$\Pr(H_p|E, I) = \frac{\Pr(E|H_p, I)\Pr(H_p|I)}{\Pr(E|I)}$$

Similarly, for proposition H_d

$$\Pr(H_d|E, I) = \frac{\Pr(E|H_d, I)\Pr(H_d|I)}{\Pr(E|I)}$$

The odds form of Bayes' theorem follows by dividing one equation by the other:

$$\frac{\Pr(H_p|E, I)}{\Pr(H_d|E, I)} = \frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)} \times \frac{\Pr(H_p|I)}{\Pr(H_d|I)}$$

More generally, if $\{H_i\}$, for $i = 1, 2, \ldots, r$, is a set of r mutually exclusive and exhaustive events or propositions, then

$$\Pr(H_i|E, I) = \frac{\Pr(E|H_i, I)\Pr(H_i|I)}{\sum_{j=1}^{r}\Pr(E|H_j, I)\Pr(H_j|I)}$$

- 3/4 of the black balls are marked H and 1/4 are marked T
- 1/4 of the white balls are marked H and 3/4 are marked T

Imagine that someone draws a ball by putting in his hand, stirring the balls, and picking one. We are asked for our probability that it is black; let's call this $\Pr(B|I)$, where I is all of the information we have been given about the container, balls, and method of drawing. It seems reasonable to assign $\Pr(B|I) = 1/3$. Now the person who drew the ball tells us that it carries the letter H. How does that affect our uncertainty about the color of the ball? What is $\Pr(B|H, I)$? Is it bigger than, the same as, or less than $\Pr(B|I)$?

If we write out the odds form of Bayes' theorem using the letters for the current example (W means white), we have

$$\frac{\Pr(B|H,I)}{\Pr(W|H,I)} = \frac{\Pr(H|B,I)}{\Pr(H|W,I)} \times \frac{\Pr(B|I)}{\Pr(W|I)}$$

The term on the left is close to what we are seeking–the odds in favor of B given that we know that the ball is marked with an H. The second term on the right can be worked out from what we started with; it is the odds in favor of B before we knew that it was marked with an H, i.e., 1/3 divided by 2/3 which is 1/2 (or two to one against). It is the first term on the right that holds the key to what we are trying to do. It is the ratio of two probabilities: the probability that a ball carries an H if it is black (we have been told that this is 3/4); and the probability that a ball carries an H if it is not black (we have been told that this is 1/4). Put another way, a ball is three times more probable to carry an H if it is black than if it is white. So our arithmetic is quite simple:

$$\frac{\Pr(B|H,I)}{\Pr(W|H,I)} = \frac{3/4}{1/4} \times \frac{1}{2} = \frac{3}{2}$$

We can easily calculate $\Pr(B|H, I)$ using the formula from the section on odds: it is 3/5.

A Model for Updating Uncertainty

We have already expressed our view that an inescapable aspect of everyday life is concerned with updating uncertainty, and hence decisions and judgments, in the light of new information. The previous example is a trivial illustration of this. We were given an initial body of information from which we assigned odds that represented our belief that the ball picked from the jar was black. It is conventional to call these the PRIOR ODDS: here the

word "prior" is used in the sense of the odds before the new information. After the information is received, we have the POSTERIOR ODDS, and these can be calculated from the prior odds using the expression in the previous section. This calculation hinges on multiplying by the ratio of two probabilities, which is called the LIKELIHOOD RATIO, and the odds form of Bayes' theorem is

$$\text{P}osterior\ odds \quad = \quad \text{L}ikelihood\ ratio \times \text{P}rior\ odds$$

Example

Here is an example to show how Bayes' theorem can be used to update one's uncertainty about something. We return to the hypothetical country where 75% of the people are Chinese and 25% are Asian Indian. A crime has been committed and the offender left a blood stain at the crime scene that has been found to be the *HUMTHO1* genotype 8,9.3. There are no eyewitnesses, but the investigator suspects that the offender was Chinese–simply because they represent 75% of the population. We have already seen that the proportion of Asian Indians with this genotype is 4.8%, but among Chinese its proportion is only 0.31%. How should this influence the investigator's view?

As before, let J denote the proposition that the offender was an Asian Indian. Then if I denotes our background information about the crime and the population of the country, the probability of J before the typing information is available is 0.25. We can write $\Pr(J|I) = 0.25$. If $\bar{J}$ denotes the event that the offender was Chinese then $\Pr(\bar{J}|I) = 0.75$. The prior odds in favor of J are

$$\frac{\Pr(J|I)}{\Pr(\bar{J}|I)} \quad = \quad \frac{0.25}{0.75} = \frac{1}{3}$$

This encapsulates the investigator's belief: the odds are 3 to 1 against the offender being Indian and hence 3 to 1 in favor of his being Chinese.

Now we have the evidence G that the offender left a stain of genotype 8,9.3. We can assign the value 0.048 to the probability of G if J is the case, and the value 0.0031 to the probability of G if $\bar{J}$ is the case. The likelihood ratio is then

$$\frac{\Pr(G|J,I)}{\Pr(G|\bar{J},I)} \quad = \quad \frac{0.046}{0.0031} \approx 15$$

So, the posterior odds are calculated as follows:

$$\frac{\Pr(J|G,I)}{\Pr(\bar{J}|G,I)} \quad = \quad \frac{\Pr(G|J,I)}{\Pr(G|\bar{J},I)} \times \frac{\Pr(J|I)}{\Pr(\bar{J}|I)} \approx 15 \times \frac{1}{3} = 5$$

The odds in favor of the offender being Asian Indian have changed by the scientific evidence from 3 to 1 against to 5 to 1 on. If we prefer to talk in terms of probability, then the probability that the offender was an Asian Indian is 5/6, or 0.83.

Exercise 1.6 There is a serious disease that affects one person in 10,000. A screening test detects the disease with a probability 0.99. However, the test can give a false positive result with probability 0.05. Let X be a person, and let I denote the information that he has been selected at random from the population. Let D denote the event that he has the disease, and E denote the event that this test gives a positive result. Calculate: (a) $\Pr(D|I)$; (b) $\Pr(E|D,I)$; (c) $\Pr(E|\bar{D},I)$; (d) The prior odds that X has the disease; (e) The likelihood ratio needed to convert prior odds of X having the disease to posterior odds; (f) The posterior odds that X has the disease.

SUMMARY

The interpretation of DNA evidence has to be made in the face of uncertainty. The origins of crime scene stains are not known with certainty, although these stains may match samples from specific people. The language of probability is designed to allow numerical statements about uncertainty, and we need to recognize that probabilities are assigned by people rather than being inherent physical quantities. As we progress through the book, it will be increasingly clear that comparing probabilities of evidence conditional on either prosecution or defense propositions is at the heart of a logical and coherent approach. The formal theory for using probabilities in this framework is provided by Bayes' theorem.

Chapter 2

Transfer Evidence

SINGLE CRIME SCENE STAIN

Imagine that a crime has been committed and an examination of the crime scene has revealed a blood stain. From its situation and apparent freshness the investigator believes that it was left by the person who committed the crime, whom we will refer to as the OFFENDER. For simplicity we will write as though offenders are male. As a result of information received–and there are several different kinds of information which could be relevant here–the investigator detains a suspect, who provides a blood sample. At the forensic science laboratory, some kind of typing technique is applied to a swab of the crime stain, which we call the CRIME SAMPLE, and the suspect's blood sample, which we call the SUSPECT SAMPLE. In keeping with the theme of this book, we suppose that it is DNA typing that has been applied. The two samples are found to be of the same type. How is this result to be interpreted? In this book we will avoid discussion of whether we should talk about genotypes (genetic constitution) or phenotype (physical type) other than to say that, for the DNA systems we have in mind, the two terms are effectively synonymous. We will therefore refer to DNA profiles as genotypes. We use the term SUSPECT although, of course, if he is later charged with the crime and the case subsequently comes to court he will become the DEFENDANT.

We assume that the scientist at the laboratory has a role in assisting the investigator, or later a court of law, in understanding the significance of this item of evidence. The investigator is interested in establishing whether there is sufficient evidence to take a case against the suspect to court. If this happens then the prosecution will endeavor to convince the jury of the suspect's guilt. We will avoid, as far as possible, talking in terms such

as GUILT and INNOCENCE. They both rest on a premise that a crime has been committed–a premise that the jury will need to consider as part of its deliberation, but which, in this case at least, is not something that is directly addressed by the DNA evidence. In the present example, the suspect may acknowledge that the blood is his and present a credible explanation why it should have been at the crime scene for purely innocent reasons. In a rape case, for example, evidence that semen on a vaginal swab from a complainant has the same type as that of the suspect does not necessarily mean that a crime has been committed. It may be that the complainant had consented to intercourse. For this reason we will talk in terms of propositions that are more closely directed to the origin of the crime stain. In the present example, the scientist can anticipate that, if the case comes to court, the prosecution will put to the jury the following proposition:

H_p: The suspect left the crime stain.

Of course, the shorthand H_p is introduced here solely to assist us in the formal analysis that follows. We would not expect prosecuting counsel to speak in algebraic notation. We have already recognized that at court the suspect will be referred to as the defendant.

We now introduce some more notation. Let G_S and G_C denote the DNA typing results for the suspect and crime sample, respectively, and let I denote the non-DNA evidence that will be put to the court in relation to H_p. Note that in the present example $G_S = G_C$. We can now view the interpretation problem as one of updating uncertainty in the light of new information. Before the DNA evidence, the probability of H_p was conditioned by I: $P(H_p|I)$. After the DNA evidence, the probability of H_p is conditioned by G_S, G_C, and I: $\Pr(H_p|G_S, G_C, I)$.

We saw in Chapter 1 how Bayes' theorem can be used to do this. However, in the present context, we cannot proceed unless we introduce some sort of alternative proposition to H_p. In the most general case we would have a range of alternatives, but things are simplest if there is only one, and in the legal setting this is appropriate because it naturally becomes the defense proposition:

H_d: Some other person left the crime stain.

Clearly, H_p and H_d in this case are mutually exclusive and exhaustive. At this point we emphasize what we will call the FIRST PRINCIPLE OF EVIDENCE INTERPRETATION: to evaluate the uncertainty of any given proposition, it is necessary to consider at least one alternative proposition.

If we now talk in terms of odds, then our problem is one of progressing from

$$\frac{\Pr(H_p|I)}{\Pr(H_d|I)}$$

which we can call the prior odds in favor of H_p, to

$$\frac{\Pr(H_p|G_S, G_C, I)}{\Pr(H_d|G_S, G_C, I)}$$

which we call the posterior odds in favor of H_p. This can be calculated from the odds ratio form of Bayes' theorem described in Chapter 1, where now $E = (G_S, G_C)$:

$$\frac{\Pr(H_p|E, I)}{\Pr(H_d|E, I)} = \frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)} \times \frac{\Pr(H_p|I)}{\Pr(H_d|I)}$$

We consider this equation to be of central importance in forensic interpretation because it enables a clear distinction to be made between the role of the scientist and that of the juror. The jury needs to address questions of the following kind:

- What is the probability $\Pr(H_p|E, I)$ that the prosecution proposition is true given the evidence?
- What is the probability $\Pr(H_d|E, I)$ that the defense proposition is true given the evidence?

On the other hand, the scientist must address completely different kinds of questions:

- What is the probability $\Pr(E|H_p, I)$ of the DNA evidence if the prosecution proposition is true?
- What is the probability $\Pr(E|H_d, I)$ of the DNA evidence if the defense proposition is true?

We cannot emphasize this distinction enough and will return to it frequently; our SECOND PRINCIPLE OF EVIDENCE INTERPRETATION is that the scientist must always be asking questions of the kind "What is the probability of the EVIDENCE given the proposition?" and not "What is the probability of the proposition given the evidence?" The latter is the kind of question that falls within the domain of the court.

The posterior odds that we seek are therefore arrived at by multiplying the prior odds by a ratio of two probabilities–the likelihood ratio (LR):

$$\mathrm{L}R \;=\; \frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)}$$

The THIRD PRINCIPLE OF EVIDENCE INTERPRETATION emerges from this analysis: The scientist must evaluate the DNA evidence, not only under the conditioning of H_p and H_d, but also under the conditioning of the non-DNA evidence, I. This is another point to which we will return frequently. But now we continue with the present example by writing out E in its two components, so the likelihood ratio is

$$\mathrm{L}R \;=\; \frac{\Pr(G_S, G_C|H_p, I)}{\Pr(G_S, G_C|H_d, I)}$$

To take this a stage further, we need to use the third law of probability to expand the numerator and denominator of the ratio. There are two ways of doing this:

$$\mathrm{L}R \;=\; \frac{\Pr(G_C|G_S, H_p, I)}{\Pr(G_C|G_S, H_d, I)} \times \frac{\Pr(G_S|H_p, I)}{\Pr(G_S|H_d, I)} \tag{2.1}$$

and

$$\mathrm{L}R \;=\; \frac{\Pr(G_S|G_C, H_p, I)}{\Pr(G_S|G_C, H_d, I)} \times \frac{\Pr(G_C|H_p, I)}{\Pr(G_C|H_d, I)} \tag{2.2}$$

The first of these is called SUSPECT ANCHORED and the second is called SCENE ANCHORED. The choice of which to use is determined by the problem at hand. In principle they should each lead to the same final answer, but one or other of them may involve probabilities that are difficult to assign given the circumstances. Interested readers are referred to Evett and Weir (1991) to see how the present problem can be solved either way. However, for the present we will use Equation 2.1.

The terms $\Pr(G_S|H_p, I)$ and $\Pr(G_S|H_d, I)$ denote the probabilities of the observation G_S for the suspect sample given either the suspect did or did not leave the crime sample. The very important point to note here is that the conditioning does not include the observation G_C for the crime sample, and whether or not the suspect left the crime sample does not provide us with any information to address our uncertainty about his genotype. So,

$$\Pr(G_S|H_p, I) \;=\; \Pr(G_S|H_d, I)$$

and the likelihood ratio simplifies to

$$LR = \frac{\Pr(G_C|G_S, H_p, I)}{\Pr(G_C|G_S, H_d, I)}$$

Now remember that $G_S = G_C$ in this particular example. If we assume that the genotype we are considering can be determined without error, then it is certain that G_C would take the value it does if H_p is the case, so $\Pr(G_C|G_S, H_p, I) = 1$ and the likelihood ratio simplifies still further to

$$LR = \frac{1}{\Pr(G_C|G_S, H_d, I)} \tag{2.3}$$

It is necessary to assign the probability of G_C given that some person other than the suspect left the crime stain. The way that we proceed from here depends on I, or, as we more commonly say, the CIRCUMSTANCES. For the present, we are going to assume that the circumstances are such that knowledge of the suspect's type G_S does not influence our uncertainty about the type of the offender, given that person is not the suspect. Then, in formal terms

$$\Pr(G_C|G_S, H_d, I) = \Pr(G_C|H_d, I) \tag{2.4}$$

Because there will be cases in which it is not true, it is important to remember that this assumption has been made. One such circumstance is the situation in which if the crime stain did not come from the suspect then it came from a close relative; then knowledge of G_S certainly influences our judgment about the probability of G_C. We will be discussing such situations in later chapters. For the time being, we assume that Equation 2.4 holds. Then

$$LR = \frac{1}{\Pr(G_C|H_d, I)} \tag{2.5}$$

Also, to simplify notation, for the rest of this section we let $G_C = G_S = G$.

We have now reached the point at which so much of the debate on DNA statistics starts; how do we assign a value to the denominator of this likelihood ratio? What is the probability that we would observe genotype G if some person other than the suspect left the stain? The answer depends entirely on the circumstances I. We need to address the concept of a group of people, identified according to the information I, to which the offender belongs, if that person is someone other than the suspect. This group will, in the simplest situation, be seen to be a POPULATION of some kind–perhaps the

population of the town in which the crime was committed, or a population of a particular ethnic group identified by some element of I, such as would be the case when an eyewitness says that the person who committed the crime appeared to be Caucasian. However the population is identified, it will not normally be the case that we know everything that there is to know about all of its members. On the contrary, our information will be limited to data collected from a small portion (a SAMPLE) of the population. The science of using samples to make inferences about populations is called STATISTICS and forms a major part of this book.

So let us assume that we have data from a sample of people we believe to be REPRESENTATIVE of the population to which the offender belongs, if the suspect is not the offender. At this stage, we do not discuss what we mean by "representative" other than to say that it is a matter for the judgment of the scientist in the case to decide whether he considers the data from the sample to be relevant to inference about the population suggested by I. Let us further assume, without going into the details of how we may do it, that we estimate that genotype G occurs in a proportion P of the population to which the offender credibly belongs, if he is not the suspect. Then we assign the probability P to the denominator of Equation 2.5 and our likelihood ratio is

$$\mathrm{L}R \;=\; \frac{1}{P}$$

If, for example, $P = 1/100$ then the likelihood ratio is 100 and the evidence could be presented in the form "The evidence is 100 times more probable if the suspect left the crime stain than if some unknown person left the stain." We discuss issues of communication later.

We wish to emphasize a couple of points here. First, in our view, the process is that the scientist ASSIGNS a numerical value to the denominator based on all the information available to him and his judgment of the relevance of all the different aspects of the information. Of course, we do not regard that judgment to be infallible and it goes without saying that, if the case comes to court, he will need to explain his reasons. Furthermore, he will have proceeded as he has by taking account of the circumstances of the offense as they were explained to him. If those circumstances change in any way, then it may be necessary for him to review his interpretation of the evidence. The scientist uses an estimate of the population proportion of the offender's type, G, as the probability of finding the crime sample to be of that type if it had been left by some other person.

Readers will note that we are emphasizing the personal nature of the interpretation and we stress that further by dismissing the idea that in any

case there is a "right answer" when it comes to the assessment of DNA evidence. Of course, there is a right answer to the question of whether or not the suspect left the crime stain but, as we have seen, it is not within the domain of the scientist's expertise to address that question (though readers familiar with the forensic field will be aware that with other types of scientific evidence, such as handwriting and fingerprinting comparison, it has long been the scientist's function to address such questions). But as far as the likelihood ratio is concerned, scientists should not be led down the false path of believing that there is some underlying precise value that is "right." We must never lose sight of the fact that, for the denominator, we are conditioning on the idea that some unknown person left the stain. What do we mean by an "unknown person"? More importantly, what will the court determine to be the most appropriate concept for the unknown person? We have loosely invoked the concept of a population, but human populations are never completely homogeneous and can never be precisely defined. Even if we become as general as possible and say that we mean the population of the world, then do we mean the population today? or yesterday? Let's say that we mean the population of the world at the instant that the crime was committed. But do we really suggest that a female octogenarian from Beijing should be regarded a potential offender in this case? Or a 6-month-old Ghanaian? If we decide that we will consider the population of the town in which the crime was committed then are we going to ignore the possibility that a person from another town left the stain? The offender may be a visitor passing through, for example. Or, do we mean an area of the town? Are some streets more likely than others to provide refuge to the true offender? Whereas we will show in Chapter 5 that such effects are in general of little practical importance, they do mean that there is always a residual uncertainty and the concept of a "right answer" is misleading. The probability that we assign to the denominator of the likelihood ratio is ultimately a matter of judgment–informed by I. This means that the scientist must be clear in his evidence to the court of the nature of I as it appeared to him when he made his assessment. If the court has a different view of I, then this will inevitably mean that the scientist must review the interpretation of the evidence.

THE BAYESIAN MODEL

Scientists have been presenting body fluid evidence in court for decades without resorting to the Bayesian model, and it is a natural reaction at this stage to suggest that presenting the evidence in the form of a likelihood

ratio is unnecessarily complicated. Why not just give the court an estimate of the frequency of the observed type?

If the case is indeed as simple as the one described above then it can be argued that the relative frequency approach is as effective as the likelihood ratio. However, as soon as any sort of complication arises–and we will meet various kinds of complications in the pages that follow–the frequency approach breaks down and can give answers that are misleading. Although we have argued that there is no "right answer" it does not follow that there are no "wrong answers": there are answers that are patently wrong–and we will meet some of them–because they lack science and logic. In cases that involve any kind of complication a Bayesian analysis is unavoidable. We claim no originality for this view, and we refer interested readers to important hallmark publications in the field. Mosteller and Wallace (1964) used Bayesian analysis to explore the issue of the authorship of "The Federalist" papers. Finkelstein and Fairley (1970) first explored transfer evidence from the perspective we have described above, and Lindley (1977) developed the ideas further in exciting ways. A comprehensive overview of evaluating scientific evidence is given in recent books by Robertson and Vignaux (1995) and Aitken (1995): the first of these is less mathematical than the second. Some description of the approach is also contained in the second NRC report (National Research Council 1996), and a very useful review was given by Friedman (1996).

Likewise, we would be remiss if we gave the impression that the Bayesian view was universally accepted–it is not. Counter views have been expressed by Tribe (1971) and Kind (1994). However, among serious students of inference relating to scientific evidence in the legal context the Bayesian view is dominant. Nevertheless, we must remember that it is a mathematical model, and any such model has limitations that we must guard against ignoring. In particular, we are going to take the evidence into the court room, where the proceedings owe no allegiance to the laws of science or mathematics and many of the participants are stubbornly nonnumerate. We will also see that there are serious problems in assigning prior probabilities. And, at the end of the day, the concept of "beyond reasonable doubt" is unquantifiable.

THREE PRINCIPLES OF INTERPRETATION

It is not our claim that Bayesian inference is a panacea for all the problems of the legal process. However, we do maintain that it is the best available model for understanding the interpretation of scientific evidence. It provides insights that are not otherwise possible. In particular, we would argue that

the preceding sections suggest three precepts for the forensic scientist:

1. To evaluate the uncertainty of any given proposition it is necessary to consider at least one alternative proposition.

   This observation is obviously more general than forensic science alone but it is well worth remembering. Within a legal trial there will be, at the very least, two competing views; one from the prosecution and one from the defense. In that situation the odds form of Bayes' theorem is applicable. There will be other situations where the hypotheses multiply, and we consider some such examples in Chapter 7.

2. Scientific interpretation is based on questions of the kind "What is the probability of the evidence given the proposition?"

   This is the most important rule to emerge from the Bayesian formulation, and we discuss it a little more in the section on the transposed conditional.

3. Scientific interpretation is conditioned not only by the competing propositions, but also by the framework of circumstances within which they are to be evaluated.

   We have demonstrated the relevance of I to the assignment of the probabilities in the likelihood ratio and we will give more examples later. It demonstrates that a scientist should make clear, either in his written report or statement or in his oral evidence, the perception of the circumstances within which the evaluation has been carried out.

ERRORS AND FALLACIES

Use of the three principles of interpretation will lead to logical statements about evidence. Failure to adhere to these principles in the past has led to some common errors and fallacies, as we now discuss.

The Transposed Conditional

It has been a long-standing practice in courts of law to indicate evidential value of a typing match between crime and suspect samples by quoting a probability based on the estimate of the proportion of a population that possesses the given type. It is conventional to make a statement of the form "the chance of observing this type if the blood came from someone other

than the suspect is 1 in 100." A common error is to reinterpret this sentence in such a form as "the chance that the blood came from someone else is 1 in 100." There are many variations on this theme; we will meet some of them later, but it is essential that all scientists who practice in this field should realize that the second of these two sentences does not follow from the first. The terminology that we have introduced in earlier sections enables us to understand the difference in the two statements.

The first sentence quoted says, in our terminology:

$$\Pr(G_C|H_d, I) = 1/100$$

The second sentence says

$$\Pr(H_d|G_C, I) = 1/100$$

The second sentence can be developed further by saying "the chance that the blood type came from someone else is 1 in 100, therefore there is a 99% chance that it came from the suspect." This makes the evidence extremely damning from the suspect's viewpoint–it seems almost certain that he left the blood stain–and so this is potentially a serious error. From the Bayesian formulation, we have seen that it is not proper for the scientist to address questions of the kind "what is the probability that the blood came from the suspect?" Such questions are the province of the court and their answer depends, not just on the DNA evidence, but also on all of the other evidence in I.

It would be pleasing to report that using a Bayesian approach removed one from the danger of transposing the conditional; alas, it is not so. We have seen that the likelihood ratio can be expressed in terms like "the evidence is 100 times more probable if the suspect left the crime stain than if some unknown person left it." Just a brief carelessness can lead to a rephrasing of this as "it is 100 times more probable that the suspect left the crime stain than some unknown person." This statement would be a reasonable statement of the posterior odds if the prior odds were exactly one: otherwise it is always wrong.

The transposed conditional is a potential trap that is always waiting for the forensic scientist. Life is made still more difficult by a very widespread tendency among members of the legal profession to commit the error in either questioning or summarizing the witnesses' evidence: "you testify that there is ... a 1 in 71 chance that a pair of contributors at random could have left the stain" (*People v. Simpson*, transcript page 33,242). The subject is discussed in more detail in Evett (1995).

Exercise 2.1 A crime has been committed, during the commission of which the offender left a bloodstain at the scene. Mr. Smith has been arrested, and a sample of his blood taken for analysis. His blood and that of the stain at the scene are found to be the same. Data from a sample of people suggest that the proportion of people in the population of potential offenders is approximately 1 in 1000. The first alternative proposition is that the stain came from Smith, and the second alternative is that the stain came from someone else. State whether each of the following methods for expressing the evidence is correct, incorrect, or ambiguous.
(a) The probability of finding this blood type if the stain had come from someone other than Smith is 1 in 1000.
(b) The probability that someone other than Smith would have this blood type is 1 in 1000.
(c) The probability that the blood came from someone other than Smith is 1 in 1000.
(d) The evidence is 1000 times more probable given the first alternative rather than the second.
(e) The first alternative is 1000 times more probable than the second.
(f) The odds are 1000 to 1 in favor of the first alternative.
(g) There is only a 1000 to 1 chance that Smith is not the donor of the bloodstain.
(h) The chance of a man other than Smith leaving blood of this type is 1 in 1000.
(i) The chance that a man other than Smith would leave blood of this type is 1 in 1000.
(j) The chance that a man other than Smith left blood of this type is 1 in 1000.
(k) It is very unlikely that the stain came from someone other than Smith.
(l) The evidence strongly supports the hypothesis that the stain came from Smith.
(Adapted from Evett 1995.)

Defense Attorney's Fallacy

The fallacy of the transposed conditional has been termed "The Prosecutor's Fallacy" (Thompson and Schumann 1987) even though defense attorneys are not immune from making the error. Another error, more likely to be made by defense attorneys, assigns prior probabilities of guilt from transfer evidence. If the DNA type of a crime stain has an estimated frequency of 1 in 100,000 people then it is true that 10 people are expected to have that type in a city of 1,000,000 people. The defense attorney's fallacy is to assume equal probabilities of guilt for these 10 people, and therefore assign a probability of guilt of 1/10 to a particular suspect from the city who does have the type. It is very doubtful that all 10 are equally likely to be guilty. We return to

this in Chapter 9.

Expected Values Implying Uniqueness

If the profile frequency is estimated to be 1 in a million, then one person in a population of a million people is expected to have that profile. When one person, the suspect, has been found in a population of this size it is fallacious to conclude that this person is guilty. Kingston (1965) discussed the issue for an example involving fingerprints.

In Chapter 3 we discuss probability distributions, but here we can point out that the actual number of people with a particular profile can be zero, one, two, or more. Each of these possible outcomes has a certain probability, and the expected value is formed by weighting each outcome by its probability. The expected value may not even be one of the possible actual values, let alone provide the only possible value. Families may have an average of 2.1 children, even though no family has that number and actual family sizes range from zero to a number considerably larger than 2.1.

When a suspect has been found to have the profile of interest, the proper question to ask is "Given that we know there is one person with this profile, what is the probability that there are other people of the same type?" We return to this when we discuss the island problem later in this chapter, and again in Chapters 4 and 9.

Defendant's Database Fallacy

The probabilities of obtaining profiles from untyped people are given numerical values on the basis of a sample of people from some population. Ideally, this sample would be drawn from among those people who could be considered possible contributors to the crime sample. This population will be defined by the hypothesis H_d and by circumstances of the crime, such as location or eyewitness reports. In practice, samples are not collected anew for every crime, and we shall see in Chapter 5 how this is accommodated. We will see there that frequencies of DNA profiles may differ between populations, especially when people in these populations have different racial backgrounds. Especially under hypothesis H_d, the racial background of the suspect does not define the population.

In 1991, a Mr. Passino was on trial for homicide in Vermont. His defense established that his paternal grandparents were Italian, his maternal grandfather was Native American and his maternal grandmother was half French and half Native American. On this basis, the defense was able to have DNA profile calculations ruled inadmissible because they were not based on a

sample of people with the same racial heritage as Mr. Passino. We pointed out (Weir and Evett 1992) the lack of logic since the defense hypothesis was that the crime sample was not from Mr. Passino, and therefore his racial background was immaterial. We agree with Lewontin (1993), who noted that the circumstances of the crime suggested that the offender may have been a member of the Abnaki tribe of Native Americans, although this still does not require account to be taken of Mr. Passino's pedigree (Weir and Evett 1993).

THE TWO-TRACE PROBLEM

We now consider a case that is a little more complicated than the first. In this case, examination of the crime scene reveals stains of two different types. Assume that the investigator is justified in inferring that the stains were left during commission of the offense, implying that there were two offenders with different blood types. Assume, further, that information received by the investigator leads to the detention of a single suspect, who provides a blood sample. DNA typing yields the genotypes G_1, G_2 from the two crime samples and G_S from the suspect sample. The suspect's type matches that of one of the stains, $G_S = G_1$, but not the other.

In this case we can visualize the two propositions as

> H_p: The suspect was one of the two men who left the crime stains.
> H_d: Two unknown men left the crime stains.

As before, we seek to evaluate

$$\mathrm{L}R \;=\; \frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)}$$

where now $E = (G_S, G_1, G_2)$. Making assumptions similar to those embodied in Equation 2.3, this can be shown to be

$$\mathrm{L}R \;=\; \frac{\Pr(G_1, G_2|G_S, H_p, I)}{\Pr(G_1, G_2|G_S, H_d, I)}$$

Remember that $G_S = G_1$.

Numerator. This is evaluated by posing the question "If the suspect and some unknown man left the crime stains, what is the probability that one of the stains would be G_1 and the other G_2?" If we repeat the assumption

that the types are determined without error, then with probability one, the suspect would leave a stain that would give genotype G_1. If we are told that the proportion of people, in what we judge to be the most relevant population, who would give observation G_2 is P_2, then the answer to the question is $1 \times P_2$, i.e.,

$$\Pr(G_1, G_2|G_S, H_p, I) \quad = \quad P_2$$

Denominator. This is evaluated by posing the question "What is the probability that two unknown people would leave stains giving observations G_1 and G_2?" Given that the proportions of the two types in the relevant population are P_1 and P_2 respectively, then it is tempting to reply, making an obvious assumption, that the answer is $P_1 \times P_2$. However, a moment's reflection is needed. Think, for a moment about tossing two coins–say a dime and a quarter–and ask yourself the probability of two heads being the result. The answer is straightforward: the dime will show a head with probability 0.5, and the quarter will also show a head with probability 0.5. If we specify that the tossing has been done properly, then these two events are independent and the probability of two heads is their product: 0.25. What now if you are asked the probability that the result of tossing the two coins is one head and one tail? There are two ways in which this can happen:

- The dime shows a head and the quarter shows a tail.
- The dime shows a tail and the quarter shows a head.

So the answer, that is another example of applying the law of total probability, is $(0.5 \times 0.5) + (0.5 \times 0.5) = 0.5$.

By exactly the same reasoning, the probability that two unknown men would give us observations G_1 and G_2 is $2P_1P_2$. You could also visualize this by imagining the two men walking in through your door: the result you seek occurs if the first man is G_1 and the second man is G_2 or if the first man is G_2 and the second man is G_1. So

$$\Pr(G_1, G_2|G_S, H_d, I) \quad = \quad 2P_1P_2$$

It follows that the likelihood ratio is

$$\mathrm{L}R \quad = \quad \frac{1}{2P_1}$$

Note that the likelihood ratio is half what it would have been if there had only been one stain. That the likelihood ratio is less in the second case is intuitively reasonable.

There is an interesting feature of this result. Note that if the proportion P_1 is greater than 0.5, the likelihood ratio is less than one. Even though there is evidence that appears to confirm the prosecution hypothesis, it actually supports the defense hypothesis. In Chapter 9, when we discuss case reporting, we will explain the unsatisfactory nature of phrases such as "consistent with." A classical forensic approach to this sort of case might have been to report a match and say something like "the evidence is consistent with the presence of the suspect's blood." We now see that this approach does not offer a balanced interpretation of the evidence.

A likelihood ratio less than one when there is a match between a suspect and the crime sample may appear counter-intuitive at first sight, but its validity can be illustrated by a simple example. Imagine that two marks have been made on a white board, one by a red pen and one by a green pen. The two pens have been dropped into an opaque container that contains 98 other pens of the same shape and size. A person is told there are now 60 red pens and 40 pens that are not red in the container, and is asked to reach into the container, withdraw a pen and speculate whether or not it was one of the two pens used to mark the white board. While the person has the pen in his hand, but before he has withdrawn his hand from the container and observed the pen, it would be reasonable for him to assign a probability of 1/50 that it was one of the two pens used. However, if he finds that the chosen pen is red, the probability that it was one of those used is 1/60. Even though the pen has the right color to have marked the white board, and so "matches," the new information (the color of the pen drawn) has actually reduced the probability that it is one of the two being sought.

TRANSFER FROM THE SCENE

In the previous two cases, the offender or offenders left evidence at the crime scene. In this case we consider interpretation when it is possible that the offender inadvertently took evidence from the crime scene.

We imagine a crime in which a victim has bled profusely after being stabbed. As a result of an investigation, a suspect is arrested and his outer clothing taken for scientific examination. Blood staining on the clothing is found to be of the same genotype G as the victim's blood. The suspect himself has a different genotype. The important difference between this and the preceding cases is that not only does the genotype match have evidential value, but also the very presence of blood staining on the suspect's clothing will need to be taken into account. We use $E = (E_1, E_2)$ to denote the two aspects of the bloodstain evidence:

- E_1: There is blood staining on the suspect's clothing.
- E_2: The blood staining on the suspect's clothing has genotype G.

For generality, let G_V and G_S denote the genotypes of the victim and suspect, respectively, remembering that $G_V = G$ and $G_S \neq G$. We consider two propositions:

H_p: The suspect is the person who stabbed the victim.
H_d: The suspect is not the person who stabbed the victim.

We will assume that the circumstances I make it clear that there was only one person involved in the assault on the victim, and there are no other mechanisms by which the victim's blood could have been transferred to the suspect. Then the likelihood ratio is

$$\mathrm{L}R = \frac{\Pr(E, G_V, G_S|H_p, I)}{\Pr(E, G_V, G_S|H_d, I)}$$

The first stages of simplification can be done using the multiplication law as follows:

$$\mathrm{L}R = \frac{\Pr(E, G_V|G_S, H_p, I)}{\Pr(E, G_V|G_S, H_d, I)} \times \frac{\Pr(G_S|H_p, I)}{\Pr(G_S|H_d, I)}$$

There is no information in H_p or H_d that would influence the genotype of the suspect, so the second ratio is one and we move to the next stage.

$$\mathrm{L}R = \frac{\Pr(E|G_V, G_S, H_p, I)}{\Pr(E|G_V, G_S, H_d, I)} \times \frac{\Pr(G_V|G_S, H_p, I)}{\Pr(G_V|G_S, H_d, I)}$$

By similar reasoning to that in the previous stage we argue that the second ratio is one. Also, if H_d is true then E is independent of G_V. Then

$$\mathrm{L}R = \frac{\Pr(E|G_V, G_S, H_p, I)}{\Pr(E|G_S, H_d, I)}$$

We will now consider the numerator and denominator in turn.

Numerator. To assist with the numerator, we introduce a new proposition T that blood was transferred from the victim to the assailant's clothing. The complementary event $\bar{T}$ is that blood was not transferred. Then we use the law of total probability:

$$\begin{aligned}\Pr(E|G_V, G_S, H_p, I) = {} & \Pr(E|T, G_V, G_S, H_p, I)\Pr(T|G_V, G_S, H_p, I) \\ & + \Pr(E|\bar{T}, G_V, G_S, H_p, I)\Pr(\bar{T}|G_V, G_S, H_p, I)\end{aligned}$$

This may look forbidding, but it can be simplified by making some reasonable assumptions:

- The probability of transfer is independent of the genotypes G_V and G_S.
- If $\bar{T}$ is the case, the blood staining evidence is independent of G_V.

The numerator of the likelihood ratio is then

$$\begin{aligned}\Pr(E|G_V, G_S, H_p, I) &= \Pr(E|T, G_V, G_S, H_p, I)\Pr(T|H_p, I) \\ &\quad + \Pr(E|\bar{T}, G_S, H_p, I)\Pr(\bar{T}|H_p, I)\end{aligned}$$

This is a situation in which the expert's judgment would, presumably, be welcomed concerning the probability that the suspect would be bloodstained if he had stabbed the victim. We will denote the expert's TRANSFER PROBABILITY as $t = \Pr(T|H_p, I)$. So

$$\begin{aligned}\Pr(E|G_V, G_S, H_p, I) &= t\Pr(E|T, G_V, G_S, H_p, I) \\ &\quad + (1-t)\Pr(E|\bar{T}, G_S, H_p, I) \qquad (2.6)\end{aligned}$$

Note that we are recognizing two explanations for the presence of blood staining: either blood was transferred during the commission of the crime or none was transferred, in which case it must have been there beforehand. In this latter case we assume that the probability of the stain evidence is the same as if the defense hypothesis is true:

$$\Pr(E|\bar{T}, G_S, H_p, I) = \Pr(E|G_S, H_d, I) \qquad (2.7)$$

Note that the right-hand term of Equation 2.7 is the same as the denominator of the likelihood ratio. We can therefore divide Equation 2.6 by this denominator and rearrange terms to get

$$LR = (1-t) + t\frac{\Pr(E|T, G_V, G_S, H_p, I)}{\Pr(E|G_S, H_d, I)} \qquad (2.8)$$

The first of these terms, being a probability, must lie between zero and one. The second term, involving a ratio of probabilities, can take values in excess of one. When this second term is greatly in excess of one, the following will not be misleading:

$$LR \approx t\frac{\Pr(E|T, G_V, G_S, H_p, I)}{\Pr(E|G_S, H_d, I)}$$

The numerator of this ratio is the answer to the question "Given that the suspect is the person who stabbed the victim and that blood was transferred, what is the probability that a stain would be found and that it has genotype

G_V, given that the victim is genotype G_V." If we assume that the genotypes are determined without error, then the answer to this question is one, so

$$\mathrm{L}R \approx \frac{t}{\Pr(E|G_S, H_d, I)} \tag{2.9}$$

Denominator. The denominator of the ratio is the answer to the question "If the suspect is not the person who stabbed the victim, what is the probability that staining with genotype G_V, different from his own type G_S, would be found on his clothing?" Evett and Buckleton (1989) showed that the question can be broken into two parts:

- What is the probability that the suspect would be found to have blood on his clothing of a type different from his own?
- What is the probability that a nonself stain on the suspect's clothing would be of type G_V?

A survey conducted by Briggs (1978) addresses the first of these questions. In a study of 122 suspects in a large murder investigation, four had clothing with blood staining in sufficient quantity to be typed and of type different from their own. Three of the four had "lightly" stained clothing and the fourth had "extensive" staining with blood from at least three sources. Clearly this small survey does not provide definitive probabilities for nonself staining, but it may guide the thinking of a forensic scientist. Write b for the probability sought in the first question.

The second question is more straightforward in that it relates to the distribution of genotypes in a population. Which population is meant? Strictly, it is the population of stains found on clothing of people belonging to the suspect's population. There are no survey data for this population. The forensic scientist is likely to have access to genotype databases constructed from samples of individuals, and this will provide the best information available. Write P_V for the estimated proportion of people with genotype G_V. The likelihood ratio in Equation 2.9 becomes

$$\mathrm{L}R \approx \frac{t}{bP_V} \tag{2.10}$$

A more rigorous derivation was given by Aitken (1995).

The apparently simple Equation 2.10 raises issues relating to the role of a forensic scientist. An extreme view would be that the scientist is little more than a technician who reports a match and provides a frequency estimate from some database. We believe that courts would look for a broader

role for forensic science. Because of his or her training and expertise, the forensic scientist must be better placed than judges or jury members to assess the evidential issues underlying the presence of nonself blood stains on a defendant's clothing. Equation 2.10 provides a framework for exercising scientific judgment, and it stimulates questions of the kind

- What is the probability that I would observe this particular pattern of blood staining if the suspect is the person who stabbed the victim?
- What is the probability that I would observe this particular pattern of blood staining if the suspect is not the person who stabbed the victim?

Even if the forensic scientist is uncomfortable in quantifying these probabilities, he or she should interpret the evidence based on the comparison of the two probabilities. The larger their ratio, the stronger is this aspect of the evidence.

THE ISLAND PROBLEM

So far, we have been concentrating on the evaluation of the LR, but now we take a broader look at the impact of transfer evidence by considering a fairly simple model, based on the idea of a closed suspect pool of known size. Several writers have studied this ISLAND PROBLEM in considerable detail (Eggleston 1983; Dawid 1994; Balding and Donnelly 1995; and Dawid and Mortera 1996).

The problem is one of a single stain at the crime scene, but now we specify the background information I in more detail. The crime was committed on a remote island where there have been no recent arrivals or departures. A suspect has been arrested because of an anonymous tip-off: evidence that cannot be put before a court. Both the genotype G_S of suspect and genotype G_C of the crime stain are the same, and the probability that a person selected at random from the island population will have this genotype is P. We also know that the island has a population of $(N + 1)$ individuals.

If the suspect is brought to court, then we consider two alternatives as before:

H_p: The suspect left the crime stain.
H_d: Some other person left the crime stain.

If there is no evidence other than the genotype match to put before the court, then disregarding the genotype evidence, he is *a priori* no more probable

than any other person on the island to have left the crime stain. We can express this by assigning the prior probability:

$$\Pr(H_p|I) = \frac{1}{N+1}$$

Robertson and Vignaux (1995) argue convincingly that this concept is compatible with the principle of the presumption of innocence. It follows that the prior odds are:

$$\frac{\Pr(H_p|I)}{\Pr(H_d|I)} = \frac{1}{N}$$

Following earlier arguments for the single crime stain case and making the same simplifying assumptions, we can say:

$$LR = \frac{\Pr(G_C|G_S, H_p, I)}{\Pr(G_C|G_S, H_d, I)} = \frac{1}{P}$$

so the posterior odds are

$$\frac{\Pr(H_p|G_S, G_C, I)}{\Pr(H_d|G_S, G_C, I)} = \frac{1}{P} \times \frac{1}{N} = \frac{1}{NP} \tag{2.11}$$

In the special case where $P = 1/(N+1)$, the expected number of people on the island with the genotype G_C on the island is 1 and the posterior odds are $(N+1)/N \approx 1$, or evens. The posterior probability, given the genotyping evidence, is $(N+1)/(2N+1) \approx 0.5$, which is in direct contrast to the fallacy we discussed earlier of implying uniqueness from an expected value of one. We return to this fallacy in Chapter 9.

Embodied in the assumptions we have made to arrive at the simple result of Equation 2.11 is the notion that the match probability P is the same for any unknown person that we consider as an alternative to the suspect. This, however, might be an unreasonable assumption, particularly if there are blood relationships between the inhabitants of the island, as we discuss in Chapter 4. It is useful, therefore, to generalize the expression to allow for varying match probabilities among the inhabitants. We will also allow for varying prior probabilities, as follows.

Let π_0 be the prior probability that the suspect left the crime stain, i.e., $Pr(H_p|I) = \pi_0$. For each of the other members of the population, let π_i, $i = 1, \ldots, N$, denote the prior probability of H_{d_i} that he or she left the crime stain. Then

$$\Pr(H_d|I) = \sum_i \Pr(H_{d_i}|I) = \sum_i \pi_i = 1 - \pi_0$$

We maintain the assumption that the probability of a match between the suspect and the crime stain, given that he left the stain, is one:

$$\Pr(G_C|G_S, H_p, I) \quad = \quad 1$$

Let P_i denote the probability that the ith other person would match, i.e., have genotype G_C:

$$\Pr(G_C|G_S, H_{d_i}, I) \quad = \quad P_i$$

Then we use the general form of Bayes' theorem, as in Box 1.4, to write out the posterior probability of H_p:

$$\Pr(H_p|G_S, G_C, I) \quad = \quad \frac{\pi_0}{\pi_0 + \sum_{i=1}^{N}(\pi_i P_i)}$$

Balding and Nichols (1995) show that it is convenient to divide the top and bottom of this expression by π_0 and to write $w_i = \pi_i/\pi_0$. Here w_i can be regarded as a weighting function that expresses how much more (or less) probable the ith person is than the suspect is to have left the crime stain, based on the non- DNA evidence. Then

$$\Pr(H_p|G_S, G_C, I) \quad = \quad \frac{1}{1 + \sum_{i=1}^{N} w_i P_i}$$

It follows that the posterior odds are

$$\frac{\Pr(H_p|G_S, G_C, I)}{\Pr(H_d|G_S, G_C, I)} \quad = \quad \frac{1}{\sum_{i=1}^{N} w_i P_i} \tag{2.12}$$

If all of the $w_i = 1$ and all of the $P_i = P$, then we have $1/NP$ as before.

The formulae for the posterior probability and odds can be used to gain some impression of the impact on the evidence of the knowledge that the suspect has a close blood relative, such as a brother, who has not been eliminated from the enquiry. For example, take P to be $1/1,000,000$ for all the inhabitants of an island of 10,000 people and assume that they all have the same prior. Then the posterior odds in favor of H_p are very close to 100 to 1 on. Now imagine that the suspect has a single brother on the island, and suppose that we calculate, by the methods we describe in Chapter 4, that the probability that a full brother would also have genotype G_C is 1/4. If we still assume that all inhabitants have the same prior then the posterior odds are approximately

$$\frac{1}{1/4 + 1/100} \quad \sim \quad 4$$

Knowledge that the suspect has a brother, who *a priori* has the same chance as anyone else on the island of being the offender, therefore has a substantial impact on the posterior odds. When P is less than 1 in a million, then the posterior odds for an island of 10,000 people are greater than 100. However, the posterior odds given the presence of a brother are still about 4. Therefore the effect of the suspect having a brother becomes greater as P becomes smaller. Note, however, that the assumption of equal prior probabilities, even for a man and his brother, let alone all inhabitants of an island, may not be at all realistic. Note, further, that the issue of a brother can be addressed directly by typing the brother.

The island problem has also been used to resolve the issue of how match probabilities should be reported in the event that a suspect has been found as a result of searching a database of DNA profiles for a profile matching an evidentiary profile (Chapter 9).

SUMMARY

The interpretation of DNA evidence rests on likelihood ratios that compare the probabilities of the evidence under alternative propositions. Although there may be occasions where simple probabilities of DNA types will be sufficient, there are many occasions where they are not. It is preferable to have a single approach for the interpretation of DNA evidence in all situations, and the one presented in this chapter will avoid fallacious statements.

Chapter 3

Basic Statistics

INTRODUCTION

We now return to the problem we faced in Chapter 2, where it was necessary for us to assign a numerical answer to the question "What is the probability that we would observe type G if some person other than the suspect left the crime stain?" Of course, it is important to understand in an individual case what we mean by the phrase "some person other than the suspect." In general, the circumstances within which the crime was committed will suggest some group of individuals to which the offender belongs, if that person is not the suspect. We will refer to that group of individuals as a POPULATION. In some situations the population might be quite tightly defined, in other cases less so. But in nearly every case we do not have information about all the members of the population. To illustrate some of the issues, we consider that the crime was committed in a hypothetical city, Gotham City, which has a populace of two million, and assume that there is no eyewitness to the crime. We are interested in the probability that some unknown person from Gotham City would give us observation G. In this book we assume that G is a genotype and that we are interested in the probability that a person has genotype G.

How are we to assign a probability to the event that an unknown member of Gotham City is genotype G? One solution would be to type every member of the population, but we can dismiss that immediately as unrealistic: even if we had the funds to do it, and even if every person consented to give a sample honestly (in real-life crime investigation screening exercises certain people may endeavor to have someone provide a sample on their behalf– Wambaugh 1989), it would take a long time, and by the time we had finished the population would have changed through births, deaths, immigration, and

emigration. In any case, are we really interested in all of the population? Men and women? Could the "unknown person" be a three year old child? Or an octogenarian? One thing is sure: whoever the unknown person is, he or she is someone with criminal tendencies since we assume that a crime has been committed. Would our time be better spent studying the genotypes of all of the criminals in Gotham City? If so, how do we do that?

So the first, and most obvious, solutions are unrealistic. In nearly all real-life situations we are going to do something else: rely on a sample of people. In the present example it would seem desirable to know the genotypes of a sample from the population of Gotham City. This is where statistics comes in. So far we have been talking just about probability and probability theory. The science of STATISTICS involves using samples to make inferences about populations.

What should we do in the context of our hypothetical case? Assuming we have agreed that it is reasonable to take a sample of the population, then how should we take the sample? The first concept we need to explore is that of a RANDOM SAMPLE. The usual definition is that this is a sample taken in such a way as to ensure that every member of Gotham City has the same chance of being selected for inclusion in the sample. In Chapter 1, we remarked that this was an alternative to the definition in which randomness was equated to uncertainty about the character (genotype) being studied. Next we must address an issue we have already touched on, that of REPRESENTATIVENESS. Remember that we are considering the idea of an unknown member of the population who committed the offense, so maybe our sample should be randomly drawn from the subset of the population that consists of people who could realistically be considered to be potential suspects for the crime (so we would presumably exclude individuals such as three-year-olds, octogenarians, and perhaps the mayor and city councillors). There are also issues of stratification. Presumably Gotham City consists of people of various ethnic backgrounds: Caucasian, Black, Hispanic, Chinese, Japanese, and so on. Should our random, representative sample be stratified to reflect the proportions of the ethnic groups (Buckleton et al. 1987)? Finally, there is the question of the sample size. This will be determined not just from statistical considerations (to be explored below in the sections on confidence intervals and Bayesian estimation) but also from considerations of cost and practicality.

Of course, a real crime laboratory would not attempt, for reasons we have touched on, to take a random, representative, stratified sample of individuals to address the question of issue. In the vast majority of cases the laboratory will have one or more CONVENIENCE samples. Such a sample may be of

laboratory staff members, or from blood donor samples with the cooperation of a local blood bank, or from samples from victims and suspects examined in the course of casework. In general, the laboratory will presumably have endeavored to collect samples from the major ethnic groups within its area of operation.

The theory we are going to describe in this section is that of drawing inferences about populations from random samples. Yet we have seen that in the forensic context, we will generally be dealing not with random, but with convenience, samples. Does this matter? The first response to that question is that every case must be treated according to the circumstances within which it has occurred, and the next response is that it is always a matter of judgment. The theory of statistics, upon which most of this book is based, operates within a framework of assumptions, but it needs to be applied to real-life problems. The forensic scientist needs to judge whether the assumptions appear reasonable in the individual case. The scientist should consider the literature, but must also ask if there is any reason to believe that knowledge of a person's sex, age, socioeconomic status, political persuasion, or tendency to criminality would in any way provide information to address the uncertainty about his genotype. In the last analysis, the scientist must also convince a court of the reasonableness of his or her inference within the circumstances as they are presented in evidence. This cause may be helped by statements in the 1996 report of the United States National Research Council (National Research Council 1996) that the loci used for identification are unlikely to be correlated with traits associated with different subsets of the population, and that frequencies of alleles at these loci do not differ very much among different subpopulations of geographic areas.

Although the term "random sample" is being applied to genotypes, not to people, we tend to ignore this distinction. Should the scientist consider his convenience sample to be a random sample? To some extent that depends on the typing system that he is considering. If it were feasible for him to identify the most relevant subset of the population of Gotham City, pick a sample at random from this subset, find the selected individuals, and persuade them to provide body fluid samples, would this truly random sample look much different from his current convenience sample? Certainly it would not be precisely the same, but would the differences have any practical effect? We believe that the scientist should be considered competent to address questions of this nature and to make judgments about what are and are not "practical effects." If the scientist is satisfied in this regard then, we maintain that he or she can proceed as though the convenience sample were a random sample.

Table 3.1: List of 50 *FES* genotypes.

10,10	10,12	11,12	12,12	10,11
10,13	11,12	10,11	11,12	10,11
11,12	8,12	10,10	10,11	10,11
10,11	10,11	10,12	10,11	11,11
11,11	10,13	12,13	10,12	11,12
10,11	11,11	10,13	11,12	11,11
11,12	10,11	11,12	11,12	11,12
10,11	11,12	10,11	11,11	11,11
10,11	10,10	10,11	11,11	10,12
10,10	10,11	11,11	8,11	10,12

Using our sample of individuals we will attempt to draw inferences about a population in order to assign a value to the denominator of the likelihood ratio. This value will be related, often in a very simple way, to the proportion of people in the population who have genotype G. But, as we have seen, we will not know that proportion, and indeed it is almost always unknowable, for the reasons we have sketched in previous paragraphs. We will use our sample, regarded as effectively random, to estimate the proportion.

Example

We will now make our example a little more concrete by saying that genotype G is 11,12 at the *HUMFES/FPS* locus, which we will call *FES* for short. Our sample consists of 50 Caucasians whose genotypes are in Table 3.1. The integers 8, 10, 11, 12, and 13 indicate alleles, and the genotype of each individual is given by the two alleles the individual possesses at this locus. One allele has come from each of the individual's parents.

We can summarize the data conveniently as the counts for all the 15 genotypes possible with 5 alleles. If previous samples had indicated the possibility of allele 9, then we may want to include those genotypes that include that allele even though they are absent for this sample. The genotype counts are displayed in the body of Table 3.2 for the alleles specified in the row and column margins.

In our sample 11 people out of 50 have the genotype 11,12. Is the figure $11/50 = 0.22$ the one that we should use to assign a value to the denominator of the likelihood ratio? The short answer is that it is the best information we have, but we must recognize that it is not the same thing as the proportion P in the population. Not only, as we have seen, is

Table 3.2: Genotype counts for the *FES* sample.

Allele	Genotype counts					
8	0					
9	0	0				
10	0	0	4			
11	1	0	15	8		
12	1	0	5	11	1	
13	0	0	3	0	1	0
Allele	8	9	10	11	12	13

the notion of population rather vague, but also a sample can provide only an ESTIMATE of the population proportion. We need to understand the properties of estimators: How good an estimator of a population proportion is a sample proportion? To understand such issues we need to study some more theory, and we next introduce the BINOMIAL DISTRIBUTION.

BINOMIAL DISTRIBUTION

An Urn Model: Two Kinds of Balls

Equal proportions of the two kinds of ball. We will learn about the binomial distribution by returning to the model in Chapter 1 of a large urn that contains a number of balls. All of the balls are indistinguishable from each other in size and shape. They differ in color, however, and for the binomial ("two names") we imagine that there are two colors: white and black. We will be considering conceptual experiments that involve drawing one ball at a time in such a way that our drawing process is completely insensitive to the color of the ball. We can do this by not looking inside the urn and by giving the balls a good stir between draws. The color we end up with after a single draw is uncertain, though we know that the greater the number of balls of a given color, the more likely we are to end up with that color. Thus, to speculate about the outcome of a draw we would like to know the proportions of the two colors in the urn. The proportions of the two colors remaining in the urn would change if we did not return the ball we drew to the urn. For this discussion we are going to return each ball we draw after we have noted its color, and this is called SAMPLING WITH REPLACEMENT.

If, before drawing a ball, we are told that half of the balls in the urn

are black and half are white, then the following statement should seem reasonable: "The probability that the ball will be black is 0.5."

The first experiment we consider consists of drawing one ball, noting its color and replacing it, and then drawing a second ball, and also noting its color. If B and W denote the colors, then there are four possible outcomes: BB, BW, WB, or WW. Each of these has a probability $0.5 \times 0.5 = 0.25$ because it seems reasonable to consider successive drawings from the urn as independent of each other. Suppose we are interested only in the number of each color, and we are indifferent to the order in which they appear. Then we can write out the three outcomes (0,1, or 2 black balls) in tabular form:

Number of black balls	Possible drawings	Number of ways	Probability
0	WW	1	0.25
1	BW, WB	2	0.50
2	BB	1	0.25

The third column reminds us that there are two ways (BW, WB–each with probability 0.25) in which we can get one black ball, but only one way in which we can get none and only one way of getting two.

We can construct a similar table for the number of black balls seen when three are drawn, noting that the probability for any particular sequence (e.g. WWB) is now $0.5 \times 0.5 \times 0.5 = (0.5)^3$

Number of black balls	Number of ways	Probability
0	1	0.125
1	3	0.375
2	3	0.375
3	1	0.125

If we wish to extend this sort of analysis to more and more balls, then we can use Pascal's triangle to work out the numbers of ways of getting a given number of blacks. The first four rows of this triangle are shown in Table 3.3, and each number in the triangle is seen to be the sum of the two numbers to the left and right of it in the preceding line. Completing the next line, corresponding to drawing six balls, should be a simple exercise for the reader.

Let's look at the fourth line in a little more detail. This is for the case in which we are drawing five balls, and we see (Table 3.4) that there are 10 ways of drawing two black balls, for example.

Pascal's triangle can be cumbersome when it comes to making such calculations for drawing large numbers of balls. To derive a general formula,

consider first that there are 5 ways of choosing an object from a set of five, 4 ways of choosing one from the remaining four, and so on. The total number of orders for five objects is $5 \times 4 \times 3 \times 2 \times 1$ and this product is written as 5! or 5 FACTORIAL. Within these $5! = 120$ ways of choosing five balls, there are 2! orders of the two black balls and 3! orders of the three white balls. These $3! \times 2! = 12$ orders do not affect the total number of black balls, so the number of ways of choosing two objects (to be black) from a list of five, termed "5 choose 2" is written as

$$ {}^5C_2 \quad or \quad \binom{5}{2} $$

and has the value

$$ \binom{5}{2} \quad = \quad \frac{5!}{2!3!} = 10 $$

More generally, the number of ways of choosing x objects from a set of n objects is

$$ \binom{n}{x} \quad = \quad \frac{n!}{x!(n-x)!} $$

Exercise 3.1 Calculate the number of ways of choosing 3 objects from 5, 5 from 12, 4 from 40, and 36 from 40.

In this model, if n balls are drawn, the probability of any particular sequence of blacks and whites is $(0.5)^n$. There are $n!/x!(n-x)!$ ways of arranging x black balls in a list of length n, so the probability of seeing a total of x black balls in n drawings is

$$ \Pr(x) \quad = \quad \frac{n!}{x!(n-x)!}(0.5)^n $$

Table 3.3: Pascal's triangle.

			1		2		1			
		1		3		3		1		
	1		4		6		4		1	
1		5		10		10		5		1

Table 3.4: Ten arrangements of three Ws and two Bs.

BBWWW	WBWBW
BWBWW	WBWWB
BWWBW	WWBBW
BWWWB	WWBWB
WBBWW	WWWBB

Exercise 3.2 When balls are drawn from an urn containing equal numbers of black balls and white balls, what is the probability of 3 out of 5 balls being black? What is the probability that 5 out of 12 are white?

Unequal proportions of the two kinds of ball. Imagine the same model, except that only 25% of the balls are black. If we draw two balls, the probability of two blacks is now $(0.25)^2$, that of two whites is $(0.75)^2$, and that of a black and a white $2 \times 0.25 \times 0.75$ (remember the two orders). The reader may like to confirm that these three probabilities sum to one. The following table lists the probabilities of the four outcomes for drawing three balls:

Number of black balls	Number of ways	Probability
0	1	$1 \times (0.75)^3 = 0.421875$
1	3	$3 \times (0.25) \times (0.75)^2 = 0.421875$
2	3	$3 \times (0.25)^2 \times (0.75) = 0.140625$
3	1	$1 \times (0.25)^3 = 0.015625$

If we were to draw n balls, then the probability that x of them would be black follows from the same kind of argument shown in this table:

$$\Pr(x) = \frac{n!}{x!(n-x)!}(0.25)^x(0.75)^{(n-x)}$$

Finally, if the proportion of black balls is p, and the proportion of white balls is $1-p$, then the probability of drawing x black balls and $n-x$ white balls is

$$\Pr(x|n,p) = \frac{n!}{x!(n-x)!}p^x(1-p)^{(n-x)} \tag{3.1}$$

and we have arrived at the BINOMIAL PROBABILITY DISTRIBUTION.

Often the binomial model is used to describe a series of independent experiments, or Bernoulli TRIALS, each of which has only two outcomes:

"success" with probability p, and "failure" with probability $q = 1 - p$. The connection with our balls-in-an-urn examples is straightforward if we regard drawing a ball as a trial, and drawing a black ball as a success.

Notation. If we have a series of n independent dichotomous trials, each with the same probability p of success, then the number of successes is a random quantity. We say it is random because we don't know its value until after the experiment (when the observed outcome becomes our DATA). Before we carry out any trials we can make a statement about the probability that the random quantity will take any particular value x. The statement is just the binomial probability in Equation 3.1. The collection of these probabilities for all $n + 1$ values of x $(0, 1, \ldots, n)$ is called the PROBABILITY DENSITY FUNCTION, or PDF. Because x can take only discrete values, the pdf is said to be a DISCRETE PDF or probability mass function. Other random quantities, such as measured lengths, can take all possible values over a range and lead to continuous distributions.

The binomial distribution, and its probability density function, are completely described by the two quantities n and p. These are the PARAMETERS of the distribution, which is written in shorthand as a $B(n, p)$ distribution.

Exercise 3.3 Five dice are rolled. What are the probabilities of two sixes, or more than two sixes?

Exercise 3.4 There are 50 unrelated people in a room. What are the probabilities that none of them have a birthday in that week, or that exactly two of them have a birthday in that week, or that more than two of them have a birthday in that week?

Exercise 3.5 Construct tables, similar to those shown for drawing two or three balls, that show the numbers of successes and the probabilities of those outcomes for: (a) The $B(6, 0.1)$ distribution; (b) The $B(6, 0.5)$ distribution. Draw a histogram for each of the distributions.

Relevance of the Binomial Model

In forensic science we do not have any particular interest in drawing balls from urns. Instead, we are interested in making inferences about allelic or genotypic proportions on the basis of samples that are typically much smaller than the populations they represent. However, we have more work to do before this model is of use to us. So far, the examples have involved making inferences about the composition of a sample, given the parameters

characteristic of the underlying population. In practice, what we are usually trying to do is estimate the one or more parameters of the population given the composition of the sample. We will return to this problem later.

Binomial Mean and Variance

It is often useful to summarize a distribution with two well-known parameters: the mean and the variance. If we sample 100 people from a population in which the proportion of genotype G is 0.2, then it seems intuitively reasonable to expect around 20 G genotypes in the sample. We wouldn't be surprised at 18 or 22, and we know that we cannot guarantee exactly 20. But 20 seems to be about the right number. This number is, in fact, the EXPECTED VALUE or MEAN.

More generally, the mean of a $B(n, p)$ pdf is np. The mean need not correspond to a value that could be obtained in a particular sample. Suppose a sample of 20 people is taken from a population in which 47% of the people support the government on some issue. Using the symbol $\mathcal{E}$ to indicate EXPECTATION, the expected number of supporters is

$$\begin{aligned}\mathcal{E}(x) &= \sum_{x=0}^{20} x \times \Pr(x|n = 20, p = 0.47) \\ &= 0 \times \Pr(x = 0|n = 20, p = 0.47) + \ldots \\ &\quad + 20 \times \Pr(x = 20|n = 20, p = 0.47) \\ &= 9.4\end{aligned}$$

Although the expected number in the sample is 9.4, this clearly cannot be the actual number in any sample. However, the sum of all possible numbers of supporters in a sample, from 0 to 20, each multiplied by its binomial probability, is 9.4. We wouldn't be surprised at 9 or 10 supporters in a sample of size 20, but we might be surprised by 18. Inference about the range of values that can be expected depends on the spread of the distribution. The best known measures of spread (or dispersion as it is more correctly called) are the VARIANCE, and its square root is known as the STANDARD DEVIATION or SD. Variance is defined as the expected value of the squared difference between values of a variable and its mean. Symbolically

$$\mathrm{V}ar(x) = \mathcal{E}(x - \mathcal{E}x)^2$$

From this definition, the variance of a $B(n, p)$ variable is $np(1 - p)$. The sd is the square root of this.

The general symbols for mean and variance are μ and σ^2, so for the binomial distribution we can write

$$\begin{aligned} \mu &= \mathcal{E}(x) = np \\ \sigma^2 &= \mathcal{E}(x-\mu)^2 = np(1-p) \end{aligned}$$

Exercise 3.6 Calculate the mean, variance and sd of a $B(400, 0.8)$ distribution.

POISSON DISTRIBUTION

The binomial distribution provides the probability of x occurrences of an event when there are n opportunities for that event to occur. For example, the number of cars that are red, out of every 10 cars that pass a certain spot on a road, might be described by this distribution. If we simply want the number of red cars that pass the spot per hour, without specifying how many cars in total there are, then the POISSON distribution may be used. Like the binomial, it gives probabilities for x occurrences of an event, but unlike the binomial the number of opportunities for the event is not specified, and need not be finite.

If the expected number of events is λ, then the Poisson probability of x events is

$$\Pr(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{3.2}$$

There is no limit to how big x may be, but it cannot be negative. The variance of the Poisson distribution is also λ, i.e.,

$$\mu = \sigma^2 = \lambda$$

The probability that the event occurs at all, e.g., that any red cars pass the spot, is the probability that x is greater than zero. This is one minus the probability that x is zero:

$$\begin{aligned} \Pr(x > 0|\lambda) &= 1 - \Pr(x = 0|\lambda) \\ &= 1 - e^{-\lambda} \end{aligned}$$

What is the corresponding result for the binomial distribution? For the binomial with parameters n, p:

$$\begin{aligned} \Pr(x > 0|p) &= 1 - \Pr(x = 0|p) \\ &= 1 - (1-p)^n \\ &\approx 1 - e^{-np}, \text{ for } large\ n \end{aligned}$$

Table 3.5: Outcomes for draws of three balls.

Number of black balls	Number of white balls	Number of red balls	Number of ways	Probability
3	0	0	1	1/27
2	1	0	3	3/27
2	0	1	3	3/27
1	2	0	3	3/27
1	1	1	6	6/27
1	0	2	3	3/27
0	3	0	1	1/27
0	2	1	3	3/27
0	1	2	3	3/27
0	0	3	1	1/27

For this reason, we can regard the Poisson as a limiting form of the binomial as n becomes large but np stays at the value λ. We will sometimes use the Poisson for very rare events, such as the occurrence of particular DNA profiles.

Exercise 3.7 In Gotham City, we have estimated that the proportion for a particular genotype G is 10^{-6}. Taking this as the true value for the proportion, and recalling that the population of Gotham City is two million, calculate the probabilities that zero, one, two, and more than two men in the city have this genotype. Assume equal numbers of men and women.

MULTINOMIAL DISTRIBUTION

An Urn Model: Three Kinds of Balls

Equal proportions of the three kinds of balls. Imagine now that the urn contains equal proportions of black, white, and red balls. Under the same conditions of drawing as before, the probability of a given color on any particular draw is 1/3. Let us start by imagining drawing three balls with replacement. There is only one way of drawing three black balls, so the probability of this is $(1/3)^3$. We can regard this as a Bernoulli trial in which black is a success and either of the other colors is a failure, so the treatment of three of a kind is simple. Other outcomes are more complicated, however, because of the wider range of possibilities, shown in Table 3.5.

Note that, when the order of balls is considered, there are a total of $3 \times$

$3 \times 3 = 27$ outcomes because each of the three balls drawn can be any of three colors. Clearly, we are dealing with a more complicated distribution than the binomial, but this is the distribution that underlies sampling whenever we have more than two types. When each of the balls has equal proportions in the urn, each of the 27 outcomes is equally likely, so the probabilities of the 10 samples (ignoring order) are just the appropriate multiples of 1/27 as shown in the table.

Unequal proportions of the three kinds of balls. When the proportions of two kinds of balls are no longer equal, the various outcomes are no longer equally probable. The same happens with more than two kinds of ball. If the proportions of black, white, and red balls in the urn are p_b, p_w and p_r, then drawing balls BWW, for example, has probability $p_b \times p_w \times p_w$.

Exercise 3.8 Using Table 3.5 as a model, calculate the probabilities for each of the 10 possible combinations of colors in a sample of three balls when the color probabilities are: black $p_b = 0.5$; white $p_w = 0.3$; red $p_r = 0.2$.

More generally, if each trial has several different outcomes, the MULTINOMIAL ("many names") distribution is appropriate. Label the different possible types resulting from each trial by i where $i = 1, 2, \ldots, m$ (m was three for the example of black, white, and red balls). Write the probability of each type at each trial by p_i, and the count of each of the types in a sample of size n as x_i. Then the probability of the set of counts $\{x_1, x_2, \ldots, x_m\}$ is

$$\Pr(x_1, x_2, \ldots, x_m) = \frac{n!}{x_1! x_2! \ldots x_m!} p_1^{x_1} p_2^{x_2} \ldots p_m^{x_m}$$

This looks formidable but it is no more than a generalization of the formula for the pdf of the binomial distribution. Indeed, the binomial pdf can be written as

$$\Pr(x_1, x_2) = \frac{n!}{x_1! x_2!} p_1^{x_1} p_2^{x_2}$$

where now x_1, x_2 are used instead of $x, n-x$. For three categories (or colors of ball in the urn model), the TRINOMIAL pdf is

$$\Pr(x_1, x_2, x_3) = \frac{n!}{x_1! x_2! x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3}$$

There is a very convenient mathematical shorthand for writing the product of similar quantities. Instead of writing $x_1!x_2!\ldots x_m!$, we can say $\prod_{i=1}^{m} x_i!$ and the multinomial pdf reduces to

$$\Pr(\{x_i\}) = \frac{n!}{\prod_{i=1}^{m} x_i!} \prod_{i=1}^{m} p_i^{x_i}$$

Fortunately, we won't have to do much work with the multinomial. When we deal with allele distributions we can often reduce the problem to considering one allele at a time and simply use the binomial distribution.

Exercise 3.9 Verify that the pdf for the trinomial gives the same results as found in Exercise 3.8.

Exercise 3.10 By combining "white, red" into a single category "not-black," verify that the trinomial formula gives the same results as does the binomial for probabilities of obtaining zero, one, two, or three black balls in a sample of three balls. Use the same probabilities as in Exercise 3.8.

NORMAL DISTRIBUTION

A $B(n, 0.5)$ distribution is symmetrical. See, for example, the bar chart for the $B(10, 0.5)$ distribution shown in Figure 3.1. The horizontal axis shows the number x of successes and the vertical axis shows the probability of x. Note that the sum of the heights of the bars is 1, reflecting the fact that it is certain that one of these 11 outcomes ($x = 0, 1, \ldots, 10$) will happen. Now look at the $B(100, 0.5)$ chart in Figure 3.2. The meanings of the two axes remain the same, but the scales have changed. However, the sum of the heights of the bars is still one.

Now look what happens when we replace the bars for $B(100, 0.5)$ with a line that joins up the midpoints of their top sides, as shown in Figure 3.3. This line looks like a smooth curve and, of course, the larger the value of n, the smoother the line becomes. The line increasingly approximates a continuous distribution called a NORMAL or GAUSSIAN distribution. If the vertical axis is scaled so that the area under the curve is one, and if x is the value along the horizontal axis, then the equation of this curve is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \tag{3.3}$$

The function $f(x)$ is the normal pdf with mean μ and variance σ^2, denoted in a shorthand notation as $N(\mu, \sigma^2)$, and its values are shown on the vertical

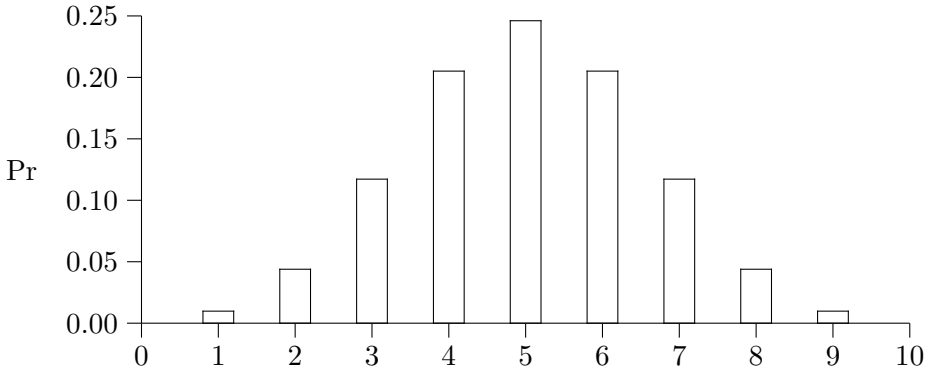


Figure 3.1: Bar chart for $B(10, 0.5)$ distribution.

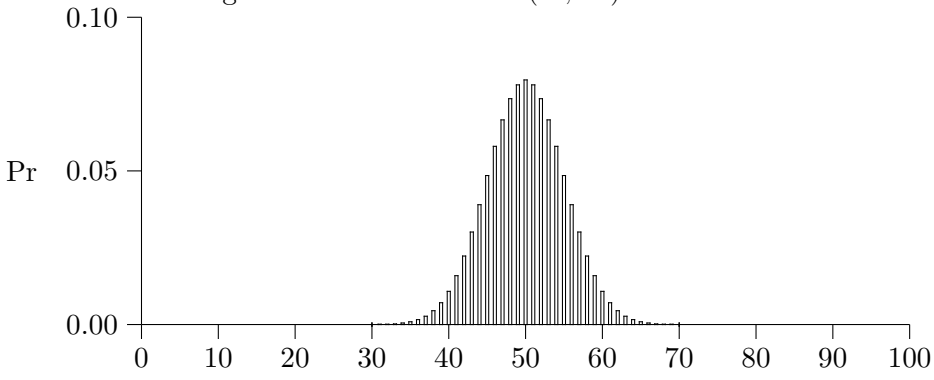


Figure 3.2: Bar chart for $B(100, 0.5)$ distribution.

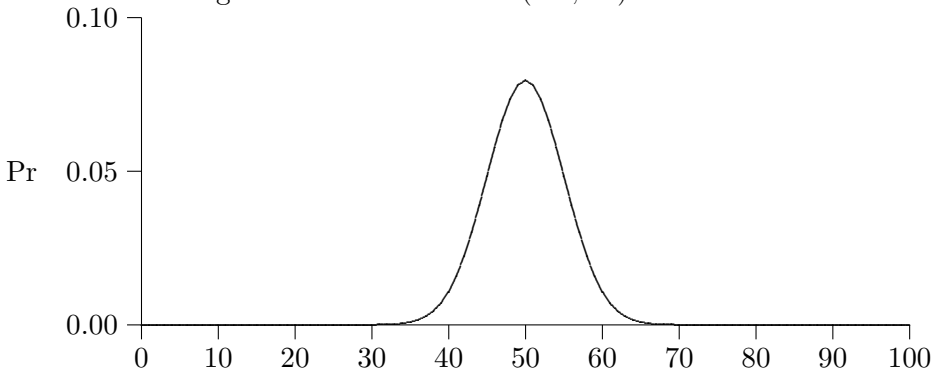


Figure 3.3: Join of midpoints of bars for $B(100, 0.5)$ distribution.

axis. As we go through the book we will encounter pdf's for unknown quantities that can take different ranges of values. As the ranges change on the horizontal scale we will see that the vertical scale ranges must also change in order to keep the area under the pdf equal to one.

The particular normal distribution approximating the binomial $B(n, p)$ has parameters $\mu = np$ and $\sigma^2 = np(1-p)$. The approximation is good for large n and for p values close to 0.5. It will not be good for very small p values because those binomial distributions are quite asymmetrical and the normal is always symmetrical.

Because the normal distribution is for continuous random quantities, such as height, the pdf gives probabilities for ranges of values, rather than for single values [$f(x)$ is not equivalent to $\Pr(x)$]. Strictly, we cannot give the probability of an unknown person in a population being exactly 1.67 m tall, but we can give the probability, for example, that the person has a height between 1.665 m and 1.675 m tall.

For a discrete quantity, the probabilities for all possible values sum to one. For a continuous quantity there are an infinite number of possible values (just as there are an infinite number of points on a line) and any one value has a zero probability. Probabilities for ranges of continuous quantities are represented by areas under the pdf between the limits for the range, and the total area under a pdf is equal to one (mathematically, the integral of the pdf over its range is one).

Equation 3.2 is not convenient for giving numerical values, and we use tables of values instead. Although there are infinitely many different normal distributions, corresponding to all the possible means and variances, they can all be rescaled into the STANDARD NORMAL that has a mean of zero and a variance of one, $N(0, 1)$. Values x of any quantity whose uncertainty is described by a $N(\mu, \sigma^2)$ distribution can be rescaled to z, with an $N(0, 1)$ distribution, by

$$z = \frac{x - \mu}{\sigma} \sim N(0, 1) \tag{3.4}$$

As we will soon see, the standard normal distribution has the useful property that about 95% of the distribution lies within two standard deviations of its mean. The distribution is described in Appendix Table A.1. Entries in the body of that table provide the probability of z being greater than the value specified in the margins. The table shows that the probability of z being greater than 1.00 is 0.1587. To find the probability of z being between 2.00 and 3.00 we use the table to find the probability (0.0228) of z being greater than 2.00 and subtract from that the probability (0.0013) of

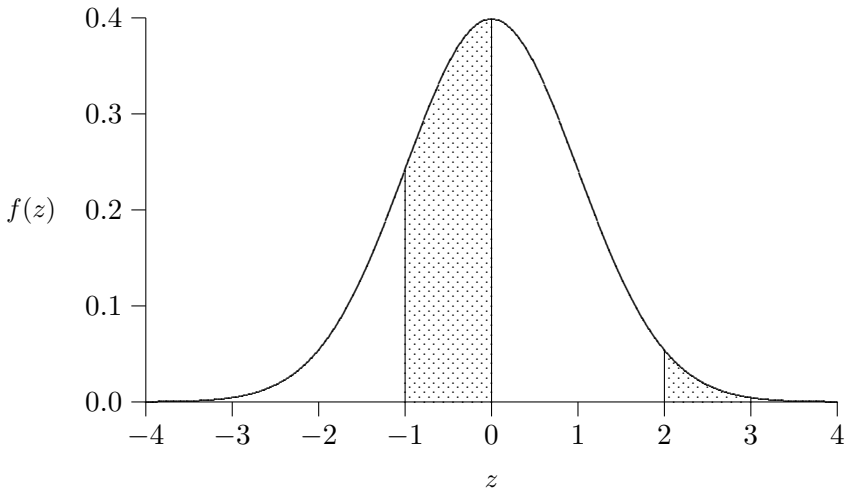


Figure 3.4: Shaded areas are probabilities of $-1 < z < 0$ and $2 < z < 3$.

z being greater than 3.00. The difference of 0.0215 is the probability for the range 2.00 to 3.00 (Figure 3.4).

Symmetry of the standard normal pdf means that the probability of z being less than -1.5, for example, is the same as that of z being greater than $+1.5$ (0.0668). Symmetry also means that the probability of z being between zero and -1.00 is the same as for the range zero to $+1.00$, and this value is the difference between the probability (0.5000) of z being greater than zero, and the probability (0.1587) of z being greater than 1.00. This difference is 0.3413 (Figure 3.4).

The most commonly used z value is 1.96. We see in Table A.1 that the area to the right of this value is 0.0250, and therefore the area to the left of 1.96 is 0.9750. This means that 1.96 is the 97.5TH PERCENTILE of the standard normal distribution. Symmetry means that a total of 5% of the area under the standard normal curve lies outside the range ± 1.96. In other words, there is a probability of 95% that a random quantity with the standard normal distribution will have a value between ± 1.96. Equation 3.4 extends this result to mean that there is 95% probability that a random quantity with any normal distribution will have a value within 1.96 standard deviations of the mean for that distribution (i.e., approximately within two standard deviations of the mean).

The standard normal approximation to the binomial variable is obtained

from

$$z = \frac{x - np}{\sqrt{np(1-p)}} \sim N(0,1)$$

and this allows probability statements to be made about counts x of a discrete binomial quantity by making use of the continuous normal distribution. For large values of n, it is much easier to refer to tables like Table A.1 than it is to evaluate $n!$.

As an example, consider sample allele proportions for the ABO blood group system. For a sample of size 16 alleles from a population in which allele A has proportion 0.50, the probabilities of x A alleles ($x = 0, 1, \ldots, 16$) are given by the $B(16, 0.5)$ distribution and are shown in Table 3.6. These values show that the probability of obtaining 6 or fewer As is 0.2272. The corresponding normal approximation to the probability is found by calculating the values of z corresponding to the range $x \leq 6$. These values are

$$\begin{aligned} z &= \frac{6 - np}{\sqrt{np(1-p)}} \\ &\leq \frac{6 - 8}{\sqrt{16 \times 0.5 \times 0.5}} \\ &= -1.0 \end{aligned}$$

Table A.1 shows that the area under the standard normal curve to the right of 1.0 is 0.1587, so the area to the left of -1.0 is also 0.1587. The normal-approximation probability of 0.1587 is somewhat different from the exact value of 0.2272. However, the quality of the normal approximation is improved with a continuity correction that reduces the magnitude of the numerator of z by 0.5. (Positive values of the numerator are reduced by 0.5 and negative values are increased by 0.5.) This makes $z = -0.75$, and the tabulated probability is then 0.2266, which is very close to the exact value.

The quality of the normal approximation to the binomial diminishes as p deviates from 0.5. For $p = 0.20$, the population proportion of the B allele, Table 3.6 shows the binomial probability of 6 or fewer Bs is 0.9733, and the corresponding z value is $z = +1.72$ (after applying the continuity correction). The tabulated normal probability is 0.9573.

INDUCTION

So far, we have considered only problems of DEDUCTION. Given a particular distribution we can make deductive statements about the outcome of a given

Table 3.6: Probabilities for $B(16, p)$ distribution.

	$p = 0.50$		$p = 0.20$	
x	$\Pr(x\|p)$	$\sum_{y=0}^{x} \Pr(y\|p)$	$\Pr(x\|p)$	$\sum_{y=0}^{x} \Pr(y\|p)$
0	0.0000	0.0000	0.0281	0.0281
1	0.0002	0.0003	0.1126	0.1407
2	0.0018	0.0021	0.2111	0.3518
3	0.0085	0.0106	0.2463	0.5981
4	0.0278	0.0384	0.2001	0.7982
5	0.0667	0.1051	0.1201	0.9183
6	0.1222	0.2272	0.0550	0.9733
7	0.1746	0.4018	0.0197	0.9930
8	0.1964	0.5982	0.0055	0.9985
9	0.1746	0.7728	0.0012	0.9998
10	0.1222	0.8949	0.0002	1.0000
11	0.0667	0.9616	0.0000	1.0000
12	0.0278	0.9894	0.0000	1.0000
13	0.0085	0.9979	0.0000	1.0000
14	0.0018	0.9997	0.0000	1.0000
15	0.0002	1.0000	0.0000	1.0000
16	0.0000	1.0000	0.0000	1.0000

experiment. We may not be able to predict the outcome with certainty but we can calculate probabilities of the various outcomes using mathematical methods that are, in principle at least, straightforward and noncontroversial. We now turn to a more difficult class of problem. Given the outcome of an experiment, how do we make inferences about the underlying distribution? For example: from a sample of n people, if we have found x occurrences of genotype G, what can we say about the proportion of genotype G in the population that has been sampled? This is the Gotham City example that we started to discuss earlier in the chapter and to which we return shortly. Here is another type of problem: From a sample of n genotypes, what can we say about whether the Hardy-Weinberg formulation (see Chapter 4) may be used for the population that has been sampled? These are examples of another kind of inference: INDUCTION. There are different schools of thought about how questions of the above type should be addressed. Leaving aside disagreement between philosophers, in the world of statistics there are the two main schools of thought that we have already referred to as FREQUENTIST and BAYESIAN. Their approaches to the solution of these problems are different, and we will attempt to explain both perspectives for the task of estimating an unknown proportion.

MAXIMUM LIKELIHOOD ESTIMATION

We return to the example in which we were attempting to assign a probability to the proposition that an unknown person in Gotham City would be genotype 11,12 at the *FES* locus. To do that we were attempting to estimate the proportion of people in the population who were that genotype (bearing in mind our discussion of the difficulties of deciding what the word population meant in this context). There was information available in a sample of 50 people (Table 3.1), 11 of whom were 11,12. Would the sample proportion $11/50 = 0.22$ serve as a good estimate of the population proportion? The short answer is yes.

We assume that the binomial distribution governs probabilities for samples from the population. If G denotes the *FES* 11,12 genotype, and p is the population proportion of this type, the probability of seeing 11 people of type G in a sample of 50 is given by the binomial $B(50, p)$ pdf:

$$\Pr(x = 11|p) \quad = \quad \frac{50!}{11!39!}(p)^{11}(1-p)^{39} \tag{3.5}$$

where we have made explicit the conditioning on p. In this form we have a function of x, based on a fixed value of p. However, another way of regarding

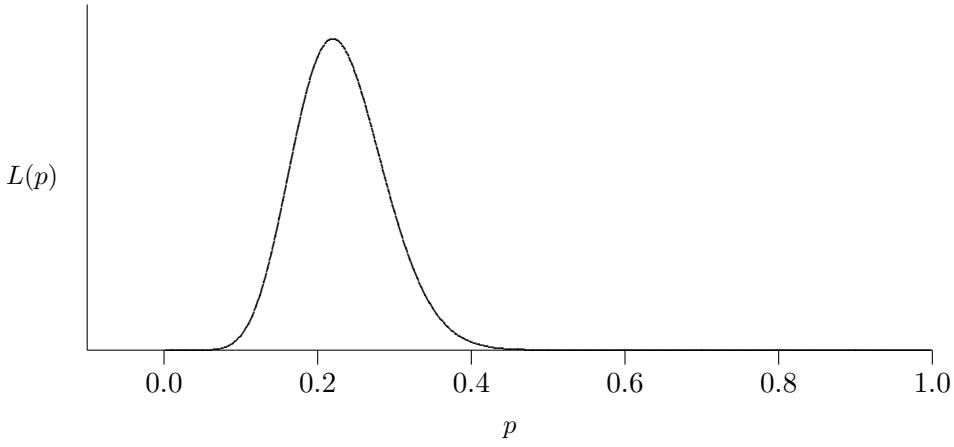


Figure 3.5: Likelihood $L(p|x = 11)$.

it is as a function of p, based on fixed value of x. We emphasize this by writing Equation 3.5 in the following way:

$$L(p|x = 11) \quad = \quad K(p)^{11}(1-p)^{39} \tag{3.6}$$

We have now defined the function of p on the right-hand side in a different way: it is called the LIKELIHOOD of p given that we know x to be 11. The likelihood is the same functional form as the pdf, up to an arbitrary constant of proportionality K. The distinction between probability and likelihood is very subtle, but also very important. We have replaced the combinatorial term in Equation 3.5 by the constant K. The shape of the likelihood curve is the same as that of the right hand side of Equation 3.5, although the vertical scale is altered. This shape is shown in Figure 3.5, where there is no vertical scale because we do not want to assign K any particular value–we are interested only in the location of the maximum of the curve. Note that, in this context, likelihood is neither a probability nor a pdf. The likelihood curve is continuous because p can take any value between 0 and 1.

Figure 3.5 has a peak at $p = 0.22$, meaning that 0.22 is the "most likely" value of the parameter given the data. The METHOD OF MAXIMUM LIKELIHOOD estimation takes this most likely value as the estimate of the parameter. The estimate is indicated by a caret, and the Box 3.1 confirms that it is given by the sample proportion

$$\hat{p} \quad = \quad \frac{11}{50} = 0.22$$

Generally, for x copies of type G in a sample of size n from a population in

Box 3.1: Maximization of likelihood
Although plotting the likelihood function is very informative, it is usually more direct to maximize the likelihood analytically. The estimate is that value that makes the first derivative zero and the second derivative negative. We write the likelihood of p as $L(p|x)$ or just L. For any count x of G genotypes,

$$L(p|x) = K(p)^x(1-p)^{n-x}$$

and it is easier to work with logarithms

$$\ln L = \ln(K) + x\ln(p) + (n-x)\ln(1-p)$$

Maximizing L is equivalent to maximizing $\ln L$. Differentiating $\ln L$ with respect to p:

$$\begin{aligned}\frac{\partial \ln L}{\partial p} &= \frac{x}{p} - \frac{n-x}{1-p} \\ \frac{\partial^2 \ln L}{\partial p^2} &= -\frac{x}{p^2} - \frac{n-x}{(1-p)^2}\end{aligned}$$

The first derivative is zero when $p = x/n$, and the second derivative is always negative. Therefore the sample proportion $\hat{p} = x/n$ is the maximum likelihood estimate.

which the proportion of type G is p,

$$\hat{p} = \frac{x}{n}$$

The maximum likelihood of a binomial (or multinomial) proportion is always the sample proportion. Why do we go to so much bother to arrive at this obvious estimate? There are more complicated situations when the parameter to be estimated is not a binomial proportion, and seeking to maximize the likelihood gives a general method of proceeding. Furthermore, maximum likelihood estimators have many desirable properties when compared to other estimators. We need to realize that the estimates themselves can take many values with differing probabilities, so that they have their own pdf's. For large samples, a maximum likelihood estimate has a normal distribution (Box 3.2), and therefore there is a 95% probability that it will be within 1.96 standard deviations of its mean. Moreover, for large samples,

Box 3.2: Asymptotic normality of maximum likelihood estimates
As the sample size becomes very large, the asymptotic normality of maximum likelihood estimates $\hat{\phi}$ of parameters ϕ can be expressed as

$$\hat{\phi} \sim N\left(\mathcal{E}(\hat{\phi}), \mathrm{V}ar(\hat{\phi})\right)$$

where $\mathcal{E}(\hat{\phi})$ is the expected value and $\mathrm{V}ar(\hat{\phi})$ is the variance of the estimate.

this mean value is just the parameter being estimated. That is, if $\hat{\phi}$ is the maximum likelihood estimate of some parameter ϕ, for large samples

$$\mathcal{E}(\hat{\phi}) = \phi$$

meaning that $\hat{\phi}$ is an unbiased estimate of ϕ. In the particular case of the binomial proportion, the result holds for all sample sizes

$$\mathcal{E}(\hat{p}) = p$$

CONFIDENCE INTERVALS

How much uncertainty is associated with population proportions estimated by maximum likelihood? In reports of public opinion surveys, we often read statements such as "47% of people surveyed support the Government, plus or minus 3 percentage points (based on a random sample of 1,000 registered voters)." This implies that 47% of the people questioned supported the Government, but that in the population as a whole the proportion is somewhere between 44% and 50%. The range (44%, 50%) is called a CONFIDENCE INTERVAL.

To see how these intervals are calculated, we return to the normal distribution. If z has the standard normal distribution, we have seen that the probability that z has a value between ± 1.96 is 0.95 (Figure 3.6). For any other quantity x that has a normal distribution with a mean of μ and a variance of σ^2, the same argument shows that with 95% probability x lies between $\mu \pm 1.96\sigma$. This is a statement about x when μ and σ are known. In practice, it is the value of x that is known and the parameter values that are unknown. Suppose first that σ is known, μ is not known, and a value of x is available. Then a similar line of reasoning leads to the interval $x \pm 1.96\sigma$ being a 95% confidence interval for μ.

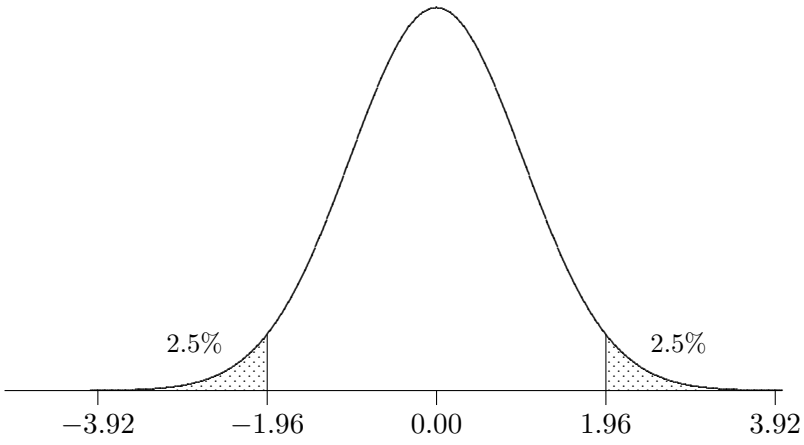


Figure 3.6: Normal distribution, showing extreme 5% of values.

The notion of confidence intervals is one belonging to frequentist statistics and it requires some care in interpretation. In the frequentist framework, probability statements cannot be made about μ because it is a fixed quantity. Instead, the interval is variable and we say there is 95% probability that the interval contains μ. In other words the procedure we follow to calculate the interval is expected to produce an interval that includes μ in 95% of the times we apply the procedure.

To provide confidence intervals for means, it is more usual to use sample means than single observations. The SAMPLE MEAN $\bar{x}$ of a sample of n observations x_i is defined as $\bar{x} = \sum_i x_i/n$, and it has a variance of σ^2/n when the variance of the distribution for the x values is σ^2. The confidence interval for μ is

$$95\%\ C.I.\ for\ \mu \quad = \quad \bar{x} \pm 1.96 s/\sqrt{n}$$

where s is the sample standard deviation for x and $s/\sqrt{n}$ is the sample standard deviation for $\bar{x}$. Strictly, use of an estimate for the standard deviation makes this an approximate 95% confidence interval. The approximation is good for large values of n; otherwise we need to use the t-distribution instead of the normal distribution. This is explained in statistics textbooks, and we will avoid the issue.

Earlier in the chapter we showed that the normal distribution can provide a good approximation to the binomial. If the probability of an event is p, the probability of x events occurring in n trials is given exactly by the binomial distribution $B(n, p)$ and approximate values of the probability can be found from the normal distribution $N(\mu, \sigma^2)$ where $\mu = np, \sigma^2 = np(1-p)$. For

the sample proportion $\hat{p} = x/n$, the normal distribution is $N(p, p(1-p)/n)$. Therefore, the 95% confidence interval for p, when $\hat{p} = x/n$ is the sample proportion, is

$$\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n} \tag{3.7}$$

What about our public opinion survey? In that example, $\hat{p} = 0.47$ and $n = 1000$. Substituting these into Equation 3.7 does indeed give a confidence interval of 0.47 ± 0.03, as suggested above.

It is important to avoid the common misconception that a confidence interval provides a probability statement about the unknown quantity. A 95% confidence interval of (0.22, 0.28) for an allele proportion, for example, should not be interpreted as meaning that there is 0.95 probability that the proportion lies in the interval. The correct interpretation is that, "in the long run," 95% of such confidence intervals will contain the population proportion. The phrase "in the long run" means that we cannot talk about this particular instance; it means that if we follow the same procedure in a large number of similar situations then the percentage of occasions in which the interval contains the correct proportion is 95%.

Exercise 3.11 For the *FES* data in Table 3.2, calculate a 95% confidence limit for the proportion of: (a) Homozygotes of type 11,11; (b) Heterozygotes of type 12,13.

BAYESIAN ESTIMATION

We now consider estimating the proportion of *FES* 11,12 genotypes in Gotham City from a Bayesian perspective. We have seen that the method of maximum likelihood gives an estimate of $11/50 = 0.22$.

As in the frequentist approach, we wish to make inferences about the unknown quantity p using the sample data. One of the features of the Bayesian approach is the recognition that we may well have some prior information about p. So let us start by reflecting on our knowledge before the sample had been collected. Let us first imagine that we had absolutely no knowledge of p. This would be a rather unrealistic state of affairs because there would have been some previous work to demonstrate that the *FES* locus was polymorphic, and that would have shed at least some light on the value, but let us discount that knowledge for the time being and imagine that

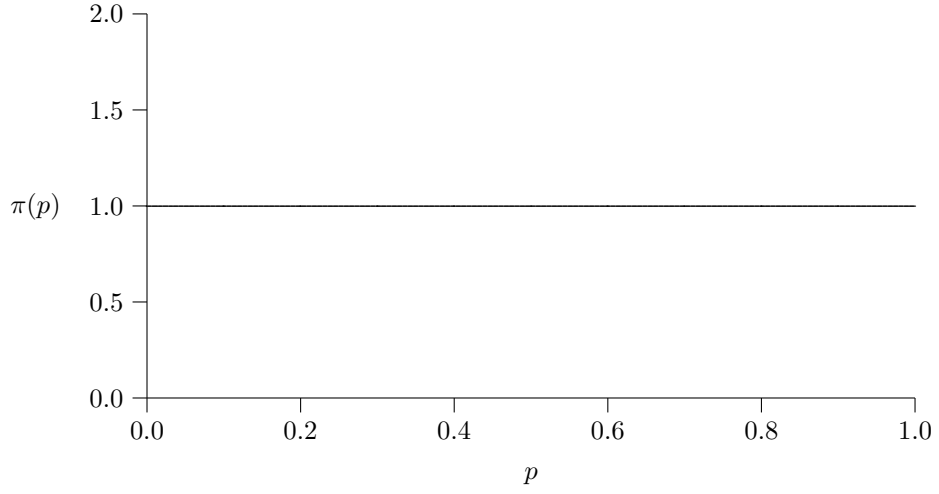


Figure 3.7: Uniform distribution.

we are unable to favor any particular value for p. One way of representing our state of knowledge employs the UNIFORM DISTRIBUTION that has the continuous pdf shown in Figure 3.7. We use the notation $\pi(p)$ for the pdf of a parameter.

Note that the value of the pdf $\pi(p)$ in Figure 3.7 is one for all values of p and this satisfies the condition that the area under the graph between $p = 0$ and $p = 1$ is one. Assigning a probability density to p is very different from the frequentist view, which does not permit probability statements about unknown parameters such as p. Using $\pi(p)$ to describe our uncertainty about the unknown parameter p is central to the Bayesian view of the problem.

Once we have a sample, we have to ask how that changes our knowledge of the pdf of p, and the solution is found in the application of Bayes' theorem. In Chapter 1, we saw how the theorem was used for weighing two hypotheses against each other:

$$\text{P}osterior\ odds \quad = \quad \text{L}ikelihood\ ratio \ \times \ Prior\ odds$$

If we are considering the probability of one of several hypotheses, then the last equation in Box 1.4 can be expressed as

$$\text{P}osterior\ probability \quad \propto \quad \text{L}ikelihood \times \ Prior\ probability \qquad (3.8)$$

When dealing with probability density functions we are essentially dealing

with an infinite number of hypotheses, and Bayes' theorem works in the same way.

Let $\pi(p)$ denote the prior pdf for p, and let $\pi(p|x)$ denote the posterior pdf. Also, let $\Pr(x|p)$ denote the probability of the data x given p. Then, for any value of p, Bayes' theorem leads to

$$\pi(p|x) \quad \propto \quad \Pr(x|p)\pi(p) \tag{3.9}$$

The term $\Pr(x|p)$, apart from a constant of proportionality, is the same as the likelihood $L(p|x)$ we met in the section on maximum likelihood estimation. It is defined by Equation 3.5 for the specific value $x = 11$, and was shown in Figure 3.5 for that x value. Now, however, instead of taking the p value that maximizes the curve as the estimator of p, we will take the additional, and important, step of combining the curve with the prior distribution as shown in Equation 3.9. The basis of the combination is simple multiplication of the two functions for every value of p. We do not describe the additional step of integration needed to provide the constant of proportionality in Equation 3.9 that ensures that the posterior pdf has an area underneath it of one. The posterior distribution is shown in Figure 3.8. In Box 3.3 we consider a more general class of prior distributions–the BETA distribution.

A few features of Figure 3.8 are worth discussing. First, note that the vertical scale runs from zero to 10 and the maximum value is about 7: this is a consequence of the requirement that the total area under of the curve be one. Note that this was not the case with the graph of the likelihood function–which is *not* a pdf. Note also that it is useful to summarize the curve by means of percentiles, which are values that divide distributions into hundreths. The MEDIAN divides a distribution in half, and 90% of a distribution lies between the 5TH and 95TH PERCENTILES.

So our posterior knowledge, based on the sample of 50, and assuming complete prior ignorance for p, can be summarized by a median value of 0.23 and 5th and 95th percentiles of 0.14 and 0.33 respectively. It would be logically legitimate for us to say that there is a 0.9 probability that p lies between 0.14 and 0.33, but we should be cautious. All of our calculations have been based on assumptions. We assumed that our sample of 50 was representative of the population relevant to answering the question posed by the denominator of the likelihood ratio; and we have also assumed that the conditions for binomial sampling have been satisfied. So we could make the probability statement about p if we wished, but we should also make the conditioning clear. The problem with this, particularly in the forensic setting, is that while we can quantify things within the framework of our assumptions, it is generally not possible for us to quantify the effects of those

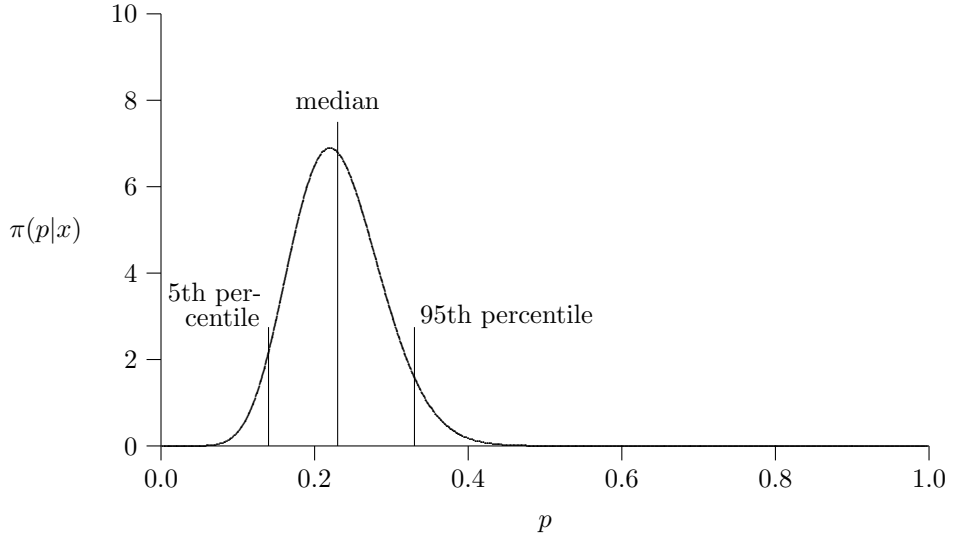


Figure 3.8: Posterior distribution for p, given $x = 11$ and the uniform prior in Figure 3.7.

Box 3.3: General Beta prior
Instead of a uniform prior, we can consider the Beta distribution $Be(\alpha, \beta)$ for p. The pdf is

$$f(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \ 0 \leq p \leq 1$$

The GAMMA FUNCTION $\Gamma(x)$ generally has to be evaluated numerically, but if x is an integer, $\Gamma(x) = (x-1)!$. When $\alpha = \beta = 1$, the Beta reduces to the uniform distribution.
Multiplying the Beta by the binomial distribution $B(2n, p)$ for a sample of $2n$ alleles, and canceling the terms not involving p gives

$$\begin{aligned} \pi(p|x) &= \frac{p^{\alpha+x-1}(1-p)^{\beta+2n-x-1}}{\int_0^1 p^{\alpha+x-1}(1-p)^{\beta+2n-x-1}dp} \\ &= \frac{\Gamma(\alpha+\beta+2n)}{\Gamma(\alpha+x)\Gamma(\beta+2n-x)} p^{\alpha+x-1}(1-p)^{\beta+2n-x-1} \end{aligned}$$

The posterior distribution is also a Beta distribution, but with parameters modified by the data. In other words, the Beta is a CONJUGATE DISTRIBUTION for the binomial. Although the whole posterior distribution is now available for p, it may be convenient to take a single feature of this distribution to serve as a Bayesian estimator of p. For example, the mean of this distribution is

$$\mathcal{E}(p|x) = \frac{\alpha+x}{\alpha+\beta+2n}$$

and the maximum of the posterior pdf is at

$$\max \pi(p|x) = \frac{\alpha+x-1}{\alpha+\beta+2n-2}$$

Table 3.7: Frequencies of genotypes in the sample of 423 British Caucasians.

8	0					
9	0	0				
10	5	0	37			
11	3	3	120	66		
12	3	0	54	66	26	
13	0	0	11	17	12	0
	8	9	10	11	12	13

assumptions and their reliability in real-world situations.

The reader may, at this stage, be rather discouraged by the size of the 90% probability interval for p but should bear in mind that we have assumed complete ignorance for our prior, and this is rarely the case in practice. We will now illustrate how this may be improved.

We may believe that the Caucasians in Gotham City are different from Caucasians elsewhere in the world, but the extensive data collected by Budowle and Monson (1993) shows that variation at restriction fragment length polymorphism (RFLP) loci is small, and studies of short tandem repeat (STR) data, though less extensive at the time of writing, suggest a similar picture. So in the light of such work, we may agree that Caucasian data collected by other workers is relevant to our problem of determining the proportion of Gotham City Caucasians who are genotype G. To illustrate how this can effect our evaluation, we take data for the *FES* locus from 423 British Caucasians collected by the Forensic Science Service as reported by Gill and Evett (1995). The data are displayed in Table 3.7.

We see that there were 66 observations of genotype 11,12 in the sample of 423. Note that the maximum likelihood estimator from this sample for p is therefore 0.16, rather than the 0.22 observed from the Gotham City sample. We could, if we thought it appropriate, use this sample to form our prior distribution for p, which would look like the curve in Figure 3.9. Note that both the horizontal and vertical scales have changed, and this marks the greater knowledge that the sample of 423 brings. The likelihood function for our data ($x = 11$) has not changed, and the posterior distribution is calculated as before. It is shown in Figure 3.10.

The posterior distribution is scarcely sharper than the prior, because the new sample is a lot smaller than the original sample, but its peak is slightly to the right, reflecting the fact that the genotype is more frequent in the Gotham City sample. The posterior median is now 0.164, and the

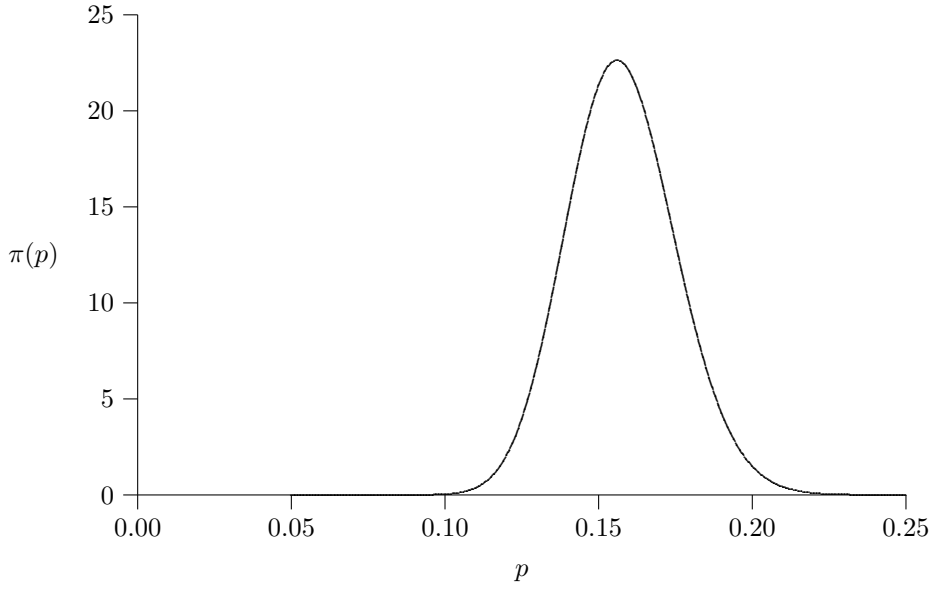


Figure 3.9: Prior distribution for p, based on a previous sample of 423 Caucasians in which 66 were of genotype 11,12.

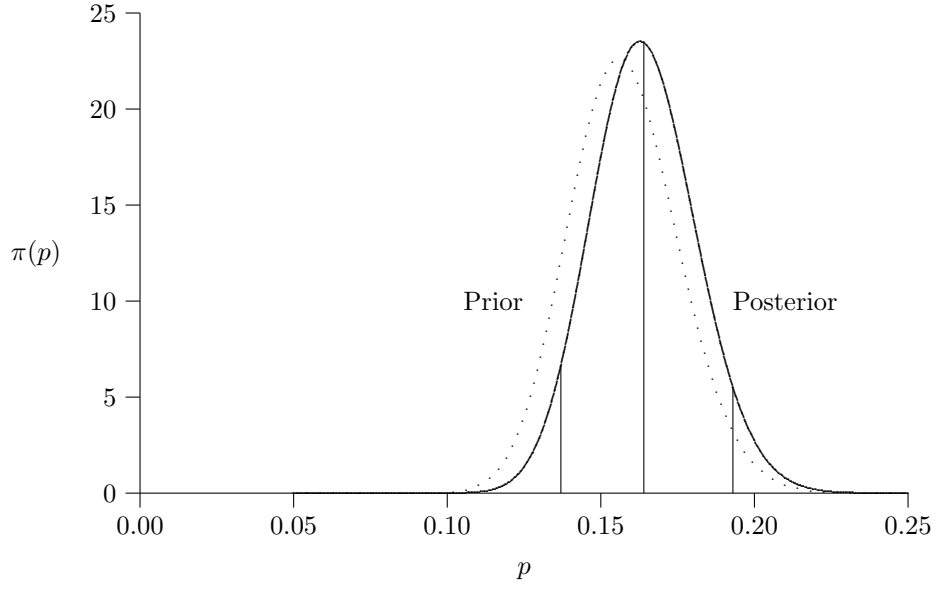


Figure 3.10: Posterior distribution (solid line) for p, given the prior distribution (dotted line) in Figure 3.9.

90% probability range is from 0.137 to 0.193.

Would this be a legitimate procedure to follow in the Gotham City example? Well, here we come once again to unquantified issues of the scientist's judgment. Is it right that the British Caucasian data should dominate the median frequency so powerfully? There is no simple answer to this, and here we must once again recognize the limit of the powers of statistics. Statistics enables quantifiable statements to be made only within a framework of assumptions; in any given situation it is the role of the scientist to make qualitative judgments.

Summary of Estimation

So far in this chapter we have concerned ourselves with problems of estimation, using as an example that of drawing inferences about the proportion of people in Gotham City who are of a certain genotype. We have seen that the method of maximum likelihood gives a single point estimate. The frequentist view also leads to the notion of a confidence interval that, in the long run, will contain the unknown value with a specified probability. We also saw that the Bayesian view of the estimation problem is to give a probability distribution for the proportion of interest. We now turn to a related issue: hypothesis testing.

TESTING HYPOTHESES

The Bayesian approach is directed to establishing a posterior probability for a hypothesis, or a posterior probability distribution for an unknown quantity. The frequentist approach is quite different in that it does not permit probability statements about hypotheses. In the same way, it is not permissible to establish a probability distribution for an unknown quantity. Instead, the frequentist approach is directed toward significance testing, the essential nature of which can be illustrated by the "goodness-of-fit" test.

Goodness-of-Fit Test

We illustrate goodness-of-fit testing by means of a simple example involving roulette wheels. Apart from the zero and double zero, roulette wheels have 36 numbers: 18 red and 18 black, and we will base our discussion on a wheel with just the 36 red or black numbers. A gambler suspects that the roulette wheel in his local casino is being operated in an unfair manner, so that red numbers come up more frequently than black numbers.

If 20 consecutive spins of the wheel result in 16 reds, is this evidence that the roulette wheel is unfair? The classical approach to this question is to set up a null hypothesis that is to be tested by the data. In this case, the null hypothesis is that the wheel is fair. For goodness-of-fit testing, the first step is to calculate the numbers of reds and blacks that would be expected if the null hypothesis were true; in this case these numbers are both 10. Next, a test statistic is devised. For categorical data, the simplest test is the CHI-SQUARE GOODNESS-OF-FIT test. This procedure compares observed and expected *counts* in *all* categories, squares the differences to remove sign, and divides by expected numbers to give greatest weight to largest proportional differences. The test statistic is written as X^2:

$$X^2 = \sum_{categories} \frac{(Observed - Expected)^2}{\mathrm{E}xpected}$$

When the null hypothesis is true, this statistic has a chi-square distribution that, in this example, has one DEGREE OF FREEDOM (DF). The degrees of freedom can be determined as the number of expected counts that can be assigned without reference to other expected counts. In this case, the expected number of reds could be set to any number from zero to 20, but then the expected number of blacks is specified. The shape of the 1 df chi-square distribution is shown in Figure 3.11, where $f(X^2)$ is the probability density for the X^2 statistic. The shaded area indicates the probability of obtaining that value of X^2, *or a greater value*, when the null hypothesis is true. These areas are displayed in Table A.2, and show that the value 3.84 delimits the largest 5% of the distribution (it is not a coincidence that 3.84 is the square of 1.96, because the square of a quantity with a standard normal distribution has a chi-square distribution with 1 df). A X^2 value greater than 3.84 would occur with probability less than 0.05 if the null hypothesis is true, so such values are used to reject the null hypothesis at the 5% SIGNIFICANCE LEVEL or with a 5% P-VALUE. Note that X^2 will be large if there are many more reds than expected, or if there are many less reds than expected. Large departures from expectation in both directions lead to rejection, and the test procedure is said to be TWO-TAILED. The term "two-tailed" refers to the hypotheses–in this case, the alternative to the hypothesis being tested has two regions: either more or less reds than expected. The term does not refer to the single tail of the chi-square distribution shown in Figure 3.11.

Does the observation of 16 red numbers in 20 spins of a roulette wheel support the hypothesis that the wheel is fair? The observed and expected counts in all categories (reds and blacks) are shown in Table 3.8, along

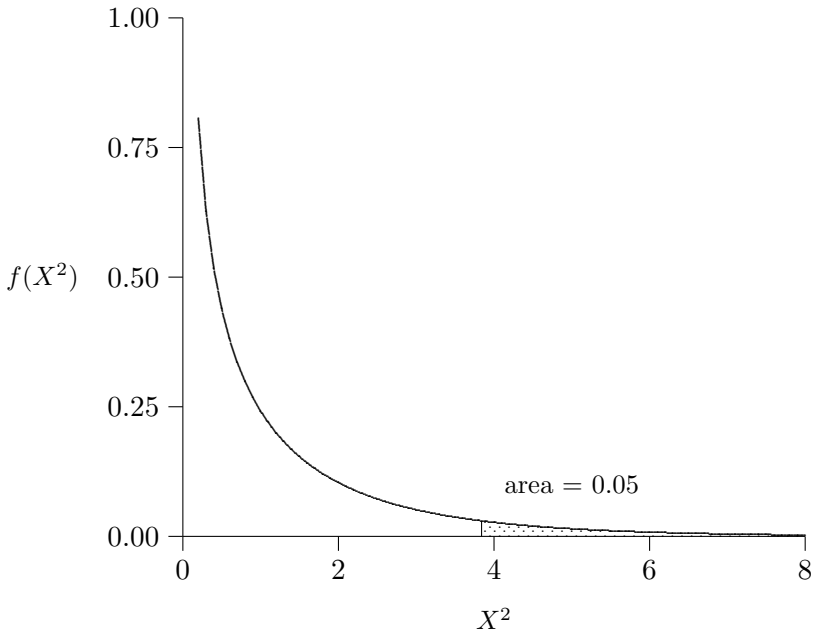


Figure 3.11: Chi-square distribution with 1 df.

with the goodness-of-fit test statistic calculations. The value of $X^2 = 7.2$ is very large compared to 3.84, and Table A.2 shows that it belongs to the least probable set of large values if the wheel is fair, where "least probable" means that set having a probability between 0.005 and 0.01. The analysis would be reported as $X^2 = 7.2(P < 0.01)$.

The goodness-of-fit test applies to more than two categories. The 36 numbers on a roulette wheel are divided into three dozens: première, milieu, and dernière. A number from each of these three is equally likely when a fair wheel is spun. What can be said about a wheel that in 20 spins gave 5, 7, and 8 numbers in the three dozens? The calculations are set out in Table 3.9. As two of the expected numbers can be assigned before the third one is automatically set, there are 2 df for the chi-square test statistic in this case, and the distribution is shown in Figure 3.12. The statistic must be greater than 5.99 to cause rejection at the 5% significance level. The calculations in Table 3.9 show that the 5, 7, 8 split is far from causing rejection. It is interesting to note that the examples in both Tables 3.8 and 3.9 follow from the same set of 20 numbers:

Table 3.8: Goodness-of-fit calculations for two categories.

Category	Observed (o)	Expected (e)	$)o-e)$	$(o-e)^2/e$
Red	16	10	6	3.6
Black	4	10	-6	3.6
Total	20	20	0	7.2

Table 3.9: Goodness-of-fit calculations for three categories.

Category	Observed (o)	Expected (e)	$(o-e)$	$(o-e)^2/e$
première	5	6.67	-1.67	0.41
milieu	7	6.67	0.33	0.01
dernière	8	6.67	1.33	0.13
Total	20	20	0	0.55

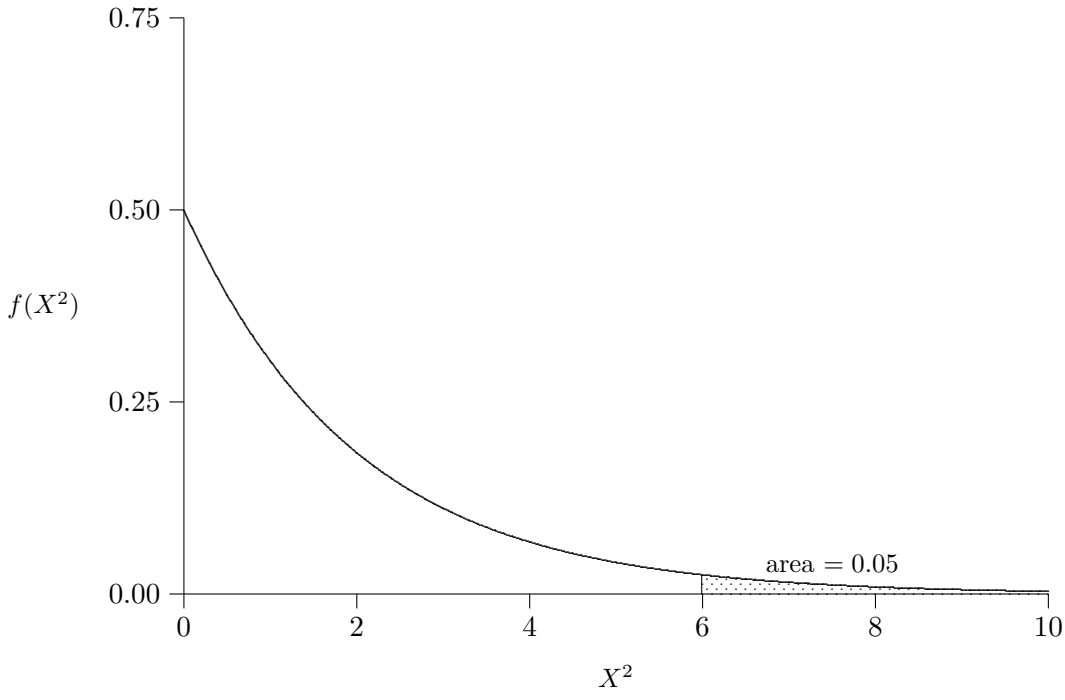


Figure 3.12: Chi-square distribution with 2 df.

22 (R)	13 (R)	32 (B)	31 (R)
16 (B)	9 (B)	25 (B)	26 (R)
33 (R)	8 (R)	28 (R)	11 (R)
17 (R)	29 (R)	20 (R)	22 (R)
2 (R)	4 (R)	29 (R)	13 (R)

Different conclusions are reached about "unbiased" by focusing on different measures of bias.

Although the chi-square goodness-of-fit test is easy to apply, it can give misleading results when expected counts are small. A category in which the expected count was 0.1 but the observed count was 1, for example, would contribute 8.1 to the test statistic and would be likely to lead to rejection of the hypothesis even though 1 is one of the two closest integers to 0.1. There have been several ad-hoc rules put forward to reduce the chance of spurious significant results, but a better procedure is to avoid the chi-square goodness-of-fit test whenever small expected counts occur. The probability tests described below offer one means of avoidance.

Exercise 3.12 A roulette wheel gave 3 black numbers in 10 spins. Would you reject the hypothesis that the wheel was fair?

Exact Test

Now that computing power is widely available, many statistical tests are being conducted as EXACT TESTS or PROBABILITY TESTS introduced by Fisher (1935). Briefly, these tests assume the hypothesis is true and calculate the probability of the observed outcome or a more extreme (less probable) outcome. Low values of this probability suggest that the hypothesis is not true.

Returning to the example of 16 reds in 20 spins of a roulette wheel, the probabilities of all 21 possible outcomes, grouped into ten pairs plus the most probable outcome, are shown in Table 3.10. We represent the number of reds by x, and because the binomial distribution is symmetrical in this case, we have chosen to group the 20 outcomes of $x \leq 9, x \geq 11$ into ten pairs: (20,0), ... , (11,9). The two outcomes in each pair have the same probability, so the probability of each pair is twice the probability of either member of the pair. The outcome of 16 reds has a probability of 0.0046 if the wheel is fair. This outcome, *or a more extreme outcome*, referring to the 10 outcomes of 20, 19, 18, 17, and 16 as well as 0, 1, 2, 3, and 4, has a 1.18% probability of occurring if the wheel is fair. Therefore the outcome of 16 belongs to the least probable 1.18% of the outcomes, and 0.0118 is called the P-value for the outcome.

The alternative to the hypothesis of fairness is two-tailed, meaning that we will reject if there are too few or two many reds. With this exact test, we are also using the two tails of the binomial distribution, unlike the use of just one tail of the chi-square distribution. A rejection region of 5% is constructed by looking for the most extreme 2.5% in each tail of the binomial. Both these tails of the binomial have outcomes a long way from the hypothesized value of x. For the chi-square distribution, only the upper tail has these extreme departures from the hypothesized value.

If we consider that 0.0118 is a low probability, then we would reject the hypothesis at the 1.18% significance level, but we acknowledge that, in so doing, we may be making an incorrect decision. The hypothesis may be true. Notice a very important feature of the frequentist view: *We cannot say there is a 1.18% chance of being wrong in this case.* What we do say is that if we imagine performing a very large number of tests like this one on roulette wheels, following the same decision rule for each test, then we would

Table 3.10: Probabilities and cumulative probabilities for $B(20, 0.5)$.

x	Probability	Cumulative probability
20 or 0	2×0.0000	0.0000
19 or 1	2×0.0000	0.0000
18 or 2	2×0.0002	0.0004
17 or 3	2×0.0011	0.0026
16 or 4	2×0.0046	0.0118
15 or 5	2×0.0148	0.0414
14 or 6	2×0.0370	0.1154
13 or 7	2×0.0739	0.2632
12 or 8	2×0.1201	0.5034
11 or 9	2×0.1602	0.8238
10	0.1762	1.0000

be in error in 1.18% of those tests in which the wheel was really true. Such errors are called TYPE I ERRORS. We return to the philosophy of hypothesis tests in Chapter 5.

A more conventional significance level is 5%, and the closest such value in Table 3.10 is 0.0414, which corresponds to rejection of the null hypothesis when x is greater than 14 or less than 6.

As another illustration of the exact test procedure, suppose a die is rolled 20 times and the quantity x is the number of times a 6 is observed. If the die is unbiased, each roll has a probability 1/6 of showing a 6, and the probabilities for each value of x follow from the binomial formula

$$\Pr(x|\mathrm{u}\textit{nbiased die}) \quad = \quad \frac{20!}{x!(20-x)!}\left(\frac{1}{6}\right)^{x}\left(\frac{5}{6}\right)^{20-x}$$

They are shown, in numerical order, in Table 3.11 along with the cumulative probability for that x or a less probable (more extreme) x. From the cumulative probability column it can be seen that the hypothesis of an unbiased die will be rejected at the 5% level if $x \geq 7$. A better statement is that for $x \geq 7$, the significance level is 0.04, and for $x = 0, x \geq 7$ the significance level is 0.06. Even though the probability for the event $x = 0$ is less than 0.05, this event does not belong to the least probable 5% of the values.

Exercise 3.13 Repeat Exercise 3.12, using an exact test.

Table 3.11: Probabilities of a 6 in 20 rolls of a fair die.

x	Prob.(x)	Cum. prob.
$\geq$12	0.0001	0.0001
11	0.0001	0.0002
10	0.0005	0.0007
9	0.0022	0.0029
8	0.0084	0.0113
7	0.0259	0.0372
0	0.0261	0.0633
6	0.0647	0.1280
1	0.1043	0.3323
5	0.1294	0.4617
2	0.1982	0.6599
4	0.2022	0.8621
3	0.2379	1.0000

Summary of Hypothesis Testing

In classical statistics, data are used to test hypotheses about specified values of some parameters. Originally this meant that a test statistic was calculated from the data and a hypothesis was either rejected or not rejected. The decision was based on comparing the calculated statistic with a tabulated set of critical values. The widespread availability of computing has brought about a move toward probability-based tests and the reporting of P-values instead of reject/not reject statements.

SUMMARY

When random samples are taken for discrete data, statistical inference rests on the binomial and multinomial distributions. Estimation of the parameters of these distributions, whether from a frequentist or a Bayesian viewpoint, rests on the likelihood function. Hypothesis testing, an aspect of frequentist inference, can be accomplished with exact tests.

Chapter 4

Population Genetics

INTRODUCTION

Until such time as DNA profiles are regarded as being sufficiently distinctive to establish individuality, forensic arguments are going to assign probabilities on the basis of proportions of marker types in a population. For biological markers that are largely determined by heritable units, this requires an understanding of population genetics.

The field of population genetics dates back to the beginning of this century, and the famous Hardy-Weinberg law was published in 1908. The field was concerned initially with the study of proportions of genes, or at least of characters controlled by genes. Understanding of what constitutes a gene has undergone many changes in recent years, and it is quite clear that most of the biological markers used for human identification should not be regarded as "genes." Nevertheless, they are dependent on portions of the human genome that are transmitted from parent to child and so fall into the domain of population genetics.

For simplicity, in this chapter we will use the term GENE to refer to any heritable unit. The alternative forms of genes are called ALLELES.

IDEAL POPULATIONS

Population genetics theory, in common with other branches of science, is based on a MODEL that incorporates a series of assumptions. Also in common with other branches of science, the models that are simple enough for tractable mathematical analysis are not true in the real world. With every model there is a compromise between simplicity and reality. For population genetics the basis of the simplest models is the concept of an IDEAL

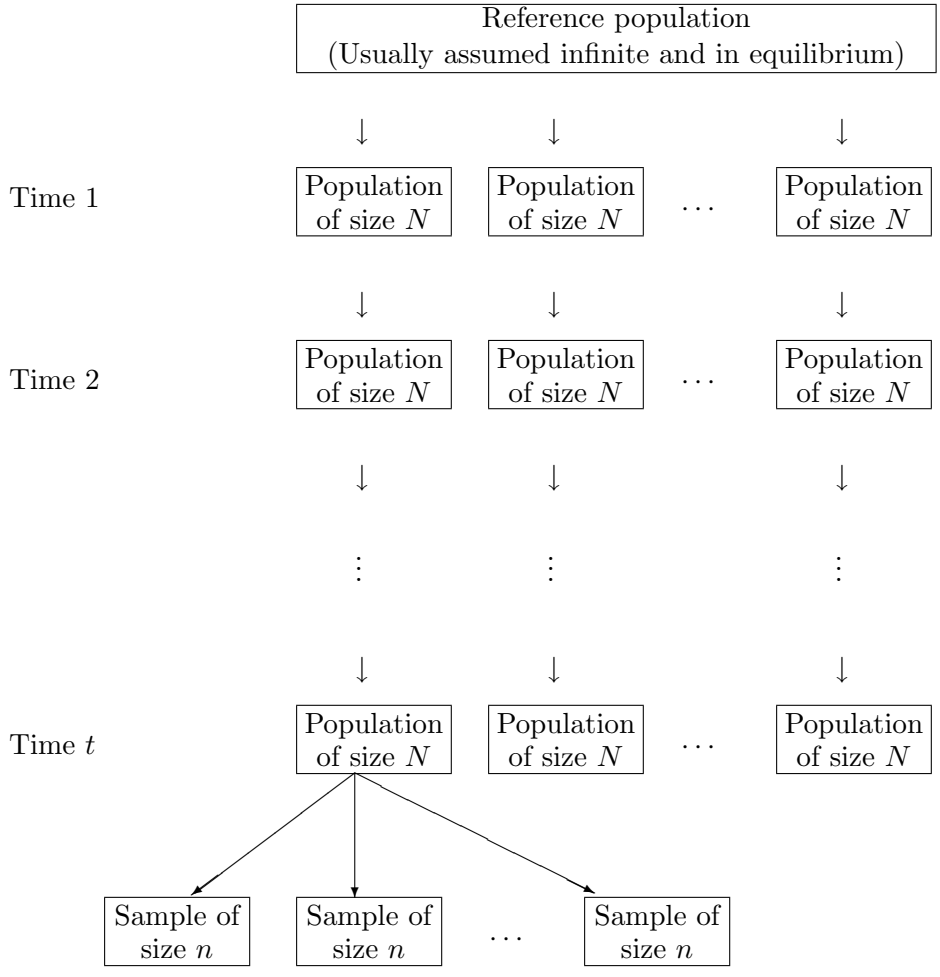


Figure 4.1: Representation of the two processes of genetic and statistical sampling.

POPULATION, and the model we will be using is summarized in Figure 4.1. The starting idea is that of a reference population from which the present population is assumed to have descended. The reference population is generally taken to incorporate the characteristics of infinite size and random mating. In this context the word "random" is used to mean that when any particular individual looks for a mate then every other member of the population has the same chance of being selected; also the mate choice of any one individual is not influenced by the mate choice of any other individual. Of course, in human populations there are two sexes but this does not affect the principles of what follows. Note that we immediately meet a contradiction because the conditions of infinite size and random mating compete with each other: pure random mating cannot occur in an infinite population because of the obvious problem that an infinite number of members of the population will be an infinite physical distance apart and so cannot possibly mate with each other! Nevertheless, this model is considered to be reliable for most real-world problems.

Next, we imagine a series of replicate populations of size N descending separately from the single reference population. In Figure 4.1 the unit of time is a single generation, which is another abstraction, of course, because in real human populations several generations coexist at any one time. These hypothetical replicate populations undergo the same evolutionary forces as the one we are studying, but differ from it because the alleles uniting to form each generation are drawn randomly from the previous generation. This random GENETIC SAMPLING makes the replicate populations different from each other. Population genetic theory has been developed to describe the variation between such replicate populations.

In practice, we rarely study complete populations; our observations are made on n individuals who have been sampled randomly from populations, and this STATISTICAL SAMPLING also gives different genetic constitutions for each sample. We have already seen that statistical theory is concerned with making inferences about populations from samples.

NOTATION

Suppose GENE **A** can take several different forms, or ALLELES, A_i. An individual can receive similar or dissimilar alleles from his or her parents. When the two alleles cannot be distinguished, so that the GENOTYPE of the individual is, for example, A_iA_i, that individual is said to be HOMOZYGOUS. Different alleles, say A_iA_j, make the individual HETEROZYGOUS.

At any particular mating, an individual transmits to the child just one of

the two alleles he has for each gene. The choice of which one is transmitted is generally considered to be random in the sense that each of the two alleles has the same probability of being transmitted. Homozygotes A_iA_i can transmit only A_i alleles, but heterozygotes can transmit either A_i or A_j with equal probability. This simple model, postulated by Mendel in 1865, forms the basis of paternity testing. If a mother of genotype A_1A_2 has a child with genotype A_2A_3 then it is clear that the child must have received an A_3 allele from its father. Men without this allele, e.g., A_1A_1, A_1A_2, or A_2A_2 men, can be excluded as being the father of that child. The genes of most use in identification are those that have very many different alleles. Whenever a suspect does not possess the alleles found in crime scene material, that suspect can be excluded from having provided the material. As another example, whenever two body parts do not share all alleles, they cannot have come from the same body. As the number of different alleles gets larger, it is less likely that two or more people will share alleles simply by chance. To quantify this statement, we need the proportions of various allele combinations in a population.

For a given population we are interested in the proportion p_i of alleles that are type A_i. This proportion is given various names by different writers; it is commonly called the ALLELE FREQUENCY, though strictly speaking, it is not a frequency but a RELATIVE FREQUENCY. We will generally use the term ALLELE PROPORTION. Likewise, we will refer to genotype proportions. The first problem that we meet in estimating p_i is that observations are usually made on genotypes rather than on single alleles. Fortunately there is a simple translation from genotypic to allele proportions in those cases where neither allele masks the appearance of the other. For such CODOMINANT alleles, both alleles present in heterozygotes can be recognized.

Codominant Allele Proportions

Proportions of codominant alleles follow from genotypic proportions by a simple counting rule. We now need to use a slightly more complicated notation for genotype proportions than was suitable for Chapters 1 to 3. We use P_{ii} to denote the proportion of A_iA_i homozygotes and P_{ij} for the proportion of A_iA_j heterozygotes. By convention, heterozygote proportions will be written with the subscripts in alphabetical or numerical order. If we are considering a gene which has three alleles, A_1, A_2, and A_3, then the three heterozygotes will be written as A_1A_2, A_1A_3, and A_2A_3 and the homozygotes as A_1A_1, A_2A_2, and A_3A_3. We will never write a heterozygote as A_2A_1, for example.

Box 4.1: Allele proportions

For alleles indexed by i, the proportion p_i for a gene with an arbitrary number of alleles, is

$$p_i = P_{ii} + \frac{1}{2}\sum_{j \neq i} P_{ij} \tag{4.1}$$

The summation term needs some explaining: $\sum_{j \neq i}$ means we sum over all values of j that differ from i. In this case we sum over all heterozygous genotypes containing allele A_i. We have said that heterozygote subscripts are written in order and this implies the additional convention that the sum includes each heterozygote only once: i.e., A_iA_j and A_jA_i do not both occur in the sum. For a locus with three alleles:

$$\begin{aligned} p_1 &= P_{11} + \frac{1}{2}(P_{12} + P_{13}) \\ p_2 &= P_{22} + \frac{1}{2}(P_{12} + P_{23}) \\ p_3 &= P_{33} + \frac{1}{2}(P_{13} + P_{23}) \end{aligned}$$

We know that all alleles in the homozygotes A_1A_1 are A_1, and half of those in the heterozygotes A_1A_2 and A_1A_3 are also A_1. So this immediately gives us an exact relationship between the allele and genotype proportions as follows:

$$p_1 = P_{11} + \frac{1}{2}P_{12} + \frac{1}{2}P_{13}$$

Note that this relationship embodies no assumptions about conditions for random mating and so on. The relation is generalized for any number of alleles in Equation 4.1 in Box 4.1.

Exercise 4.1 Find the proportions of the three alleles from the following set of genotypic proportions:

A_1A_1	A_1A_2	A_2A_2	A_1A_3	A_2A_3	A_3A_3
0.36	0.36	0.09	0.12	0.06	0.01

Dominant Allele Proportions

DOMINANT alleles prevent their RECESSIVE counterparts from being observed, and this makes the translation of genotype proportions to allele proportions difficult. One of the most common examples is provided by the ABO blood group system. Alleles A and B are both dominant to O, and this results in only four recognizable PHENOTYPES from the six possible genotypes:

Genotype	AA	AO	BB	BO	AB	OO
Phenotype	A	A	B	B	AB	O

Simple counting procedures cannot now provide the allele proportions from the phenotype proportions. The counting procedure would require knowledge of the six genotype proportions.

RANDOM MATING

As we have seen, the simplest assumptions in population genetics are those that are clearly not true. Chief among these is that populations are infinite in size and mate at random, but the consequences of this theory are so powerful that we explore them in some detail. One of the consequences is that knowledge of the genotype of one member of a mating pair provides no information about the genotype of the other member of the pair: this is what we have called INDEPENDENCE. If a person with genotype A_iA_j is selected from a population in which such types have proportion P_{ij}, then a second person drawn at random from the population has the same chance of having this genotype. The proportion did not change after selecting the first person, as it would if the genotypes were dependent or if the population were finite.

What are the consequences of mating at random between members of an infinite population? To see this, consider a gene with two alleles A_1 and A_2, with proportions p_1 and p_2, respectively, and three genotypes A_1A_1, A_1A_2, and A_2A_2. There are nine possible mate pairs, each of which has a certain probability of having the three offspring genotypes, as shown in Table 4.1.

It is the assumption of random mating, and the consequent independence, that allows the probability of each mate pair to be written in the third column as the product of the two separate genotype probabilities. The last three columns show, for each parental combination, the probability of each of the three possibilities for the children of that combination.

Now we introduce the term P'_{11} as the proportion of the A_1A_1 genotypes among the children, i.e., the second generation. We can calculate P'_{11} from

Table 4.1: Demonstration of Hardy-Weinberg law.

			Children		
Mother	*Father*	*Probability*	A_1A_1	A_1A_2	A_2A_2
A_1A_1	A_1A_1	$P_{11} \times P_{11}$	1	0	0
	A_1A_2	$P_{11} \times P_{12}$	1/2	1/2	0
	A_2A_2	$P_{11} \times P_{22}$	0	1	0
A_1A_2	A_1A_1	$P_{12} \times P_{11}$	1/2	1/2	0
	A_1A_2	$P_{12} \times P_{12}$	1/4	1/2	1/4
	A_2A_2	$P_{12} \times P_{22}$	0	1/2	1/2
A_2A_2	A_1A_1	$P_{22} \times P_{11}$	0	1	0
	A_1A_2	$P_{22} \times P_{12}$	0	1/2	1/2
	A_2A_2	$P_{22} \times P_{22}$	0	0	1

the terms in rows 1, 2, 4, and 5 of the table, using the law of total probability, as follows:

$$\begin{aligned} P'_{11} &= (1)P_{11}^2 + (1/2)P_{11}P_{12} + (1/2)P_{12}P_{11} + (1/4)P_{12}^2 \\ &= [P_{11} + (1/2)P_{12}]^2 \end{aligned}$$

Now we can apply Equation 4.1, which tells us that the expression in the brackets is simply p_1, so

$$P'_{11} = p_1^2$$

If we define P'_{12} as the proportion of A_1A_2 children, then from rows 2 through 8 in Table 4.1

$$\begin{aligned} P'_{12} &= (1/2)P_{11}P_{12} + (1)P_{11}P_{22} + (1/2)P_{12}P_{11} + (1/2)P_{12}^2 \\ &\quad + (1/2)P_{12}P_{22} + (1)P_{22}P_{11} + (1/2)P_{22}P_{12} \\ &= 2[P_{11} + (1/2)P_{12}][P_{22} + (1/2)P_{12}] \\ &= 2p_1p_2 \end{aligned}$$

Similarly we can show that the proportion of A_2A_2 children is $P'_{22} = p_2^2$. This shows that the offspring genotype proportions are specified completely by parental allele proportions as

$$P'_{11} = (p_1)^2, \;\; P'_{12} = 2p_1p_2, \;\; P'_{22} = (p_2)^2 \tag{4.2}$$

This is a demonstration of the HARDY-WEINBERG LAW. We can now use Equation 4.2 to calculate p'_1, the proportion of allele A_1 among the children:

$$\begin{aligned} p'_1 &= P'_{11} + \frac{1}{2}P'_{12} \\ &= p_1^2 + p_1 p_2 \\ &= p_1(p_1 + p_2) \\ &= p_1 \end{aligned}$$

The allele proportions are unchanged in the second generation. Note that the assumption of random mating has led to the Hardy-Weinberg law for the genotypes of the children, even though we made no assumption about relationships between the genotypic and allele proportions in the parental generation. We did invoke the counting rule of Equation 4.1, but that always holds for codominant alleles. For the more general case of an unspecified number of alleles the Hardy-Weinberg law is

$$\left.\begin{aligned} P_{ii} &= p_i^2 \\ P_{ij} &= 2p_i p_j, \quad j \neq i \end{aligned}\right\} \tag{4.3}$$

Note that we now have two separate ways of relating allelic and genotypic proportions: Equations 4.1 and 4.3. As we have seen, Equation 4.1 is always true and codominant allelic proportions can always be found from genotypic proportions in this way. Equation 4.3 enables genotypic proportions to be found as products of allelic proportions, but this should generally be regarded as only an approximation. The Hardy-Weinberg law was demonstrated above under the assumptions of random mating in an infinite population, without other forces such as selection, mutation, or migration. Under these circumstances, the law holds in all generations after the first, and so describes an EQUILIBRIUM situation. As will be shown in the next section, however, the law may hold even if these circumstances do not occur, i.e., it can hold when there is selection (Lewontin and Cockerham 1959) or when there is nonrandom mating (Li 1988).

Exercise 4.2 Assuming Hardy-Weinberg equilibrium, find the three genotypic proportions for a gene with allele proportions of 0.7 and 0.3 for A_1 and A_2.

Exercise 4.3 Assume Hardy-Weinberg equilibrium to find the proportions of the four blood group types, A, B, AB, and O when the allele proportions are $p_A = 0.2, p_B = 0.1$, and $p_O = 0.7$.

DISTURBING FORCES

In the previous section we showed that, in infinite random-mating populations, allele proportions do not change from the parent to offspring generations and, if the conditions for Hardy-Weinberg equilibrium apply, then they will remain constant through all generations. Proportions can change if there are disturbing forces, such as selection, mutation, and migration.

Selection

Selection refers to the differential abilities of genotypes to contribute to the next generation. One mode of selection is VIABILITY SELECTION, which involves the ability of an individual to survive to adulthood. If these abilities depend on the genotype of an individual, allele proportions will be altered. To illustrate the kinds of arguments that can be made, suppose the three genotypes A_1A_1, A_1A_2, and A_2A_2 at locus **A** have VIABILITIES w_{11}, w_{12}, and w_{22}. This changes the genotype proportions from P_{11}, P_{12}, P_{22} at the beginning of the generation to P'_{11}, P'_{12}, and P'_{22} at the end of the generation, where

$$\begin{aligned} P'_{11} &= w_{11}P_{11}/\bar{w} \\ P'_{12} &= w_{12}P_{12}/\bar{w} \\ P'_{22} &= w_{22}P_{22}/\bar{w} \end{aligned}$$

Dividing by the MEAN VIABILITY $\bar{w}$ ensures that the proportions still add to one:

$$\bar{w} = w_{11}P_{11} + w_{12}P_{12} + w_{22}P_{22}$$

Allele proportions will change over time until an equilibrium is established. The equilibrium may reflect the loss of an unfavorable allele: p_1 will become zero if $w_{22} > w_{12} > w_{11}$ for example. In the special case of HETEROZYGOTE ADVANTAGE, $w_{11} < w_{12} > w_{22}$, it can be shown (Box 4.2) that a polymorphic equilibrium will be established. This is a situation where heterozygotes are the "fittest" and will always remain in the population. As long as they remain, it is guaranteed that both alleles will also remain in the population.

Lewontin and Cockerham (1959) showed that if $w_{11}w_{22} = w_{12}^2$ then the Hardy-Weinberg relation also holds after selection:

$$P'_{11} = (p'_1)^2, \;\; P'_{12} = 2p'_1p_2, \;\; P'_{22} = (p'_2)^2$$

Box 4.2: Allele proportion changes under selection

The allele proportions after selection are

$$\begin{aligned} p_1' &= P_{11}' + (P_{12}'/2) \\ &= [w_{11}P_{11} + (w_{12}P_{12}/2)]/\bar{w} \\ p_2' &= [w_{22}P_{22} + (w_{12}P_{12}/2)]/\bar{w} \end{aligned}$$

and if the original genotype proportions obey the Hardy-Weinberg law, $P_{11} = p_1^2, P_{12} = 2p_1p_2, P_{22} = p_2^2$,

$$\begin{aligned} p_1' &= p_1[w_{11}p_1 + w_{12}p_2]/\bar{w} \\ p_2' &= p_2[w_{22}p_2 + w_{12}p_2]/\bar{w} \end{aligned}$$

If an equilibrium is established, there is no change in proportion, $p_1' = p_1$, and the equilibrium value is written as $\hat{p}_1$. Then

$$\bar{w} = w_{11}\hat{p}_1 + w_{12}\hat{p}_2 = w_{22}\hat{p}_2 + w_{12}\hat{p}_2$$

and this can be rearranged to provide

$$\left.\begin{aligned} \hat{p}_1 &= \frac{(w_{12} - w_{22})}{(w_{12} - w_{11}) + (w_{12} - w_{22})} \\ \hat{p}_2 &= \frac{(w_{12} - w_{11})}{(w_{12} - w_{11}) + (w_{12} - w_{22})} \end{aligned}\right\} \tag{4.4}$$

These equilibrium allele proportions are valid, lying between zero and one, for heterozygote advantage: $(w_{12} - w_{11}) > 0, (w_{12} - w_{22}) > 0$.

Although this is a very restrictive situation, it makes the point that consistency with Hardy-Weinberg proportions does not mean that selection is absent.

In spite of this elegant theory, there are very few cases known where human genes exhibit these kinds of single-gene selective forces. The one case that recurs in textbooks is that of sickle-cell anemia. Sickle-cell hemoglobin is produced by a single substitution in the DNA sequence for β-hemoglobin. If A represents the normal form and S represents the sickle form, then AA individuals are normal. Affected homozygotes SS suffer from a severe hemolytic anemia, and their reproductive fitness is very low under primitive living conditions. The AS heterozygotes are clinically healthy, and it appears that they are at a reproductive advantage in the primitive conditions because they are protected against Falciparum malaria. Some studies (Allison 1954) have estimated the following selection coefficients, measured relative to that for heterozygotes:

$$\begin{aligned} w_{AA} &= 0.7961 \\ w_{AS} &= 1.0000 \\ w_{SS} &= 0.1698 \end{aligned}$$

Substituting these values into Equation 4.4 in Box 4.2 provides

$$\hat{p}_S = \frac{0.2039}{0.8302 + 0.2039} = 0.1972$$

This value has some support in empirical studies.

For some human diseases, one homozygote has a fitness of essentially zero. With cystic fibrosis, for example, it has been rare for victims to survive beyond childhood. If the disease allele is D and the normal form A, then this means that $w_{DD} = 0$. It is the DD people who have the disease. For the disease allele to be maintained in the population, a possible mechanism is heterozygote advantage. Setting the fitness of the heterozygote AD to 1, this translates into a selective disadvantage for the AA homozygote, $w_{AA} = 1 - s$, where s is a positive quantity less than one, and Equation 4.4 gives

$$\hat{p}_D = \frac{s}{1+s}$$

The proportion of affected people born in the population will be

$$\hat{P}_{DD} = \hat{p}_D^2$$

Some studies in mice suggest that the cystic fibrosis alleles confer protection against cholera. Other possibilities have also been suggested. The resulting advantage to individuals carrying the disease allele may be sufficient to have maintained the allele in the population.

Box 4.3: Allele proportion changes under mutation to and from the allele

The proportion of the A_1 allele will change as follows:

$$p_1' = (1-\mu)p_1 + \nu p_2$$

The first term shows the reduction in proportion caused by alleles mutating from A_1 to A_2, and the second term shows the increase in proportion caused by alleles mutating from A_2 to A_1. Likewise

$$p_2' = \mu p_1 + (1-\nu)p_2$$

Note that in the offspring generation the two proportions still sum to one.
At equilibrium $\hat{p}_1 = p_1' = p_1$, and $\hat{p}_2 = p_2' = p_2$. Making this substitution provides

$$\left.\begin{aligned} \hat{p}_1 &= \frac{\nu}{\mu+\nu} \\ \hat{p}_2 &= \frac{\mu}{\mu+\nu} \end{aligned}\right\} \tag{4.5}$$

Mutation

Selection can change the proportions only of existing alleles. Human genetic diversity is caused, ultimately, by the introduction of new alleles from mutation. A convenient definition of mutation refers to the event that a child carries an allele different from those in its parents. In the present context, the change is spontaneous rather than being caused by some mutagen.

A simple mathematical model for a locus with two alleles A_1 and A_2 assigns the probability μ that any A_1 allele will mutate to the other allele A_2 during transmission from parent to child. Likewise, ν is the probability that A_2 will mutate to A_1. The probabilities μ and ν are called MUTATION RATES. This might be an appropriate model for RESTRICTION SITES, which can be lost by a single base change but may need several base changes to be created from a random sequence. The mutation rates toward and away from restriction sites are different. The consequences of these opposing mutation rates are shown in Box 4.3, and eventually the population reaches an equilibrium with allele frequencies given in Box 4.3 by Equation 4.5.

Another combination of events leading to a polymorphic equilibrium would be the continued introduction of an allele by mutation and selection against that allele. As an example, suppose that A_2 is detrimental. This

can be expressed by assigning fitnesses as follows:

$$w_{11} = 1, w_{12} = 1 - hs, w_{22} = 1 - s$$

The quantity h is sometimes called the degree of DOMINANCE. Then the proportion of A_2 will become small over time but it will not be lost from the population if there is continuing mutation, at rate μ, to that allele. We show in Box 4.4 that an equilibrium proportion is reached with the approximate value

$$\hat{p}_2 \approx \begin{cases} \sqrt{\frac{\mu}{s}}, & h = 0 \\ \frac{\mu}{hs}, & h \neq 0 \end{cases}$$

Mutation rates are generally quite small, and are unlikely to be as high as 0.01. The selection coefficients can also be quite small: if $h = 0$ and s is four times the mutation rate, the allele frequency p_1 will eventually become 0.5. There is no way that the magnitudes of selection and mutation rates can be deduced simply from allele frequencies in the current population, and estimation of very small coefficients will be difficult without large sample sizes.

A quite different population genetic model of mutation supposes that every mutation results in a new type of allele, as might be appropriate when alleles refer to long stretches of DNA. This is the INFINITE ALLELES model. If mutation were the only force acting, each allele would quickly become very rare. However, there is a process, GENETIC DRIFT, which can counter the increase in genetic variation brought about by mutations, and this will be discussed later.

Migration

Alleles do not occur in the same proportions in different populations. For the ABO blood group gene, for example, allelic proportions in some different countries are shown in Table 4.2. Differences between populations will tend to diminish over time as a consequence of GENE FLOW. This may come about because of migration of people between populations, or because of marriage between people from different populations.

One migration model, which may be appropriate for an idealized version of the United States Caucasian population, is called the ISLAND MODEL. The population consists initially of a series of separate subpopulations, in this case descended from different European countries. The model has a

Box 4.4: Allele proportion changes under selection and mutation

From Box 4.2, selection changes the proportion of A_1 from p_1 to p_1^* according to

$$p_1^* = p_1[p_1 + (1 - hs)p_2]/\bar{w}$$

Because the A_2 allele will be in low proportion, there will be a very low contribution of mutation from A_2 to A_1. Ignoring this contribution:

$$\begin{aligned} p_1' &\approx (1-\mu)p_1^* \\ &= p_1(1 - hsp_2)(1-\mu)/\bar{w} \end{aligned}$$

At equilibrium, $p_1' = p_1 = \hat{p}_1$, and this equation becomes

$$(1 - hs\hat{p}_2)(1-\mu) = \bar{w} = 1 - 2hs\hat{p}_2(1-\hat{p}_2) - s\hat{p}_2^2$$

If allele A_2 is recessive, $h = 0$, and

$$s\hat{p}_2^2 = \mu$$

but if $h \neq 0$ and $\hat{p}_2$ is considered to be small enough that $\hat{p}_2^2 \approx 0$

$$hs\hat{p}_2(1+\mu) \approx \mu$$

and the $1 + \mu$ term can be approximated by 1 to give the result in the text.

Table 4.2: ABO allele proportions in samples from different countries.

Country	p_A	p_B	p_O
Argentina	0.024	0.015	0.961
Australia	0.333	0.000	0.667
Austria	0.294	0.107	0.599
Bolivia	0.028	0.004	0.967
Burundi	0.192	0.119	0.689
Chad	0.113	0.144	0.744
Chekoslovakia	0.205	0.222	0.573
Chile	0.056	0.003	0.941
Costa Rica	0.001	0.001	0.998
Egypt	0.222	0.104	0.674
Haiti	0.125	0.180	0.695
Hungary	0.297	0.141	0.562
Iraq	0.230	0.156	0.605
Jordan	0.220	0.146	0.634
Malaysia	0.178	0.203	0.619
Samoa	0.258	0.130	0.610
Scotland	0.214	0.070	0.716
Solomon Is.	0.274	0.058	0.668
Spain	0.292	0.065	0.642

Source: Roychoudhury and Nei (1988).

Box 4.5: Allele proportion changes under island-model migration

The equation for change in allele proportion is rearranged as

$$p' = (1-m)p + [1-(1-m)]\bar{p}$$

If we write p'' for the proportion in the second generation, then the same equation gives

$$\begin{aligned} p'' &= (1-m)p' + [1-(1-m)]\bar{p} \\ &= (1-m)^2 p + [1-(1-m)^2]\bar{p} \end{aligned}$$

If we continue this process, at generation t we find that the proportion in the subpopulation becomes p^t:

$$p^t = (1-m)^t p + [1-(1-m)^t]\bar{p}$$

As t becomes large, $(1-m)^t$ approaches zero, so ultimately the proportion becomes simply $\bar{p}$.

proportion m of the alleles in each subpopulation migrating out each generation, and an equal proportion m migrating in from the rest of the entire population. If $\bar{p}$ is the average proportion of allele A in the entire population, and if p is the proportion in a particular subpopulation in some generation, then in the next generation

$$p' = p - mp + m\bar{p}$$

where the first term represents the initial proportion, the second represents a loss of that allele by migration away from the subpopulation, and the third represents migration into the subpopulation from the rest of the population. This is really for an "infinite island" model, so that $\bar{p}$ is the same whether or not the particular subpopulation is included in the average. The quantity m is the migration rate. In the absence of any other forces, migration will keep the average proportion $\bar{p}$ constant over time, and each subpopulation p will move toward this average value. This can be thought of as a redistribution of all the A alleles so that they become equally represented in every subpopulation. Details are shown in Box 4.5.

Exercise 4.4 If a population consists of five equal-sized subpopulations, with proportions for some allele A of 0.1, 0.3, 0.5, 0.7, and 0.9, what are the proportions of the allele in the generation after the initial one when the migration rate is 10%

per generation?

The gene flow process appropriate for the (idealized) US African-American population may be described by another model: that of ADMIXTURE. There has been interbreeding between Caucasian and African-American populations to the point that the present African-American population has a different genetic constitution from that in the African populations of origin. Suppose the proportion of allele A in the ancestral African population was p_a, and that a fraction m of the alleles in the present day admixed population came from the Caucasian population. If the allele proportion in the Caucasian population is p_c, then the proportion p_m in the admixed population is

$$p_m = (1-m)p_a + mp_c$$

so that

$$m = \frac{p_m - p_a}{p_c - p_a}$$

Vogel and Motulsky (1986) give an example for the fy^a allele of the Duffy blood group system. In the present–day African-American population of Oakland, California the allele proportion is 0.0941, in western African populations it is essentially zero, and in the Oakland Caucasian population it is 0.4286. Hence

$$m = \frac{0.0941 - 0}{0.4286 - 0} = 0.2195$$

This figure of about 20% has also been found for alleles in the *ABO* blood group system. Chakraborty, Kamboh, et al. (1992) found a figure of about 25% for some data from Pittsburgh.

Heterogeneous Populations

A concept related to migration and admixture is that of population heterogeneity. Regardless of intermarriage or migration, allele and genotypic proportions can be calculated for each subpopulation separately or for the combined population. When allele proportions are different in the subpopulations, there may appear to be Hardy-Weinberg disequilibrium in the population as a whole even if there is equilibrium in each subpopulation. This phenomenon is known as the WAHLUND EFFECT and could arise if each subpopulation had random mating but very little gene exchange with other subpopulations. The principle can be illustrated with a simple example.

Suppose a population consists of two equal-sized subpopulations with proportions of alleles A, B, and C at locus *HBGG*:

	Subpopulation		Total
Allele	1	2	population
A	0.6	0.4	0.5
B	0.3	0.1	0.2
C	0.1	0.5	0.3

The proportions shown for the total population are the averages of those in the two subpopulations. Assuming Hardy-Weinberg proportions in each subpopulation, the genotypic proportions are now shown. Once again, the total population values are the averages of those in the two subpopulations.

	Subpopulation		Total	
Genotype	1	2	population	(HW)
AA	0.36	0.16	0.26	(0.25)
BB	0.09	0.01	0.05	(0.04)
CC	0.01	0.25	0.13	(0.09)
AB	0.36	0.08	0.22	(0.20)
AC	0.12	0.40	0.26	(0.30)
BC	0.06	0.10	0.08	(0.12)

The last column of this table shows, in parentheses, the genotypic proportions expected if the total population were in Hardy-Weinberg equilibrium. These are obtained by multiplying the total allele proportions. Note that all homozygotes have higher proportions than would be predicted by the Hardy-Weinberg law. No general statement can be made for heterozygotes, although on average they will have lower proportions than predicted. They can, however, have higher proportions, as is the case for AB. It would be wrong, for example, to change homozygote proportions to allow for heterogeneous populations but not to change heterozygote proportions as was suggested by the 1996 NRC report (National Research Council 1996) in its Recommendation 4.1. A general treatment is given in Box 4.6.

Exercise 4.5 Suppose a population consists of proportions 0.5, 0.3, and 0.2 of three subpopulations, each of which is in Hardy-Weinberg equilibrium for gene **A**, but with different proportions, 0.4, 0.6, and 0.2 respectively, for allele A. What are the proportions of the A allele and the AA genotype in the combined population?

INBREEDING

We now begin to discuss population genetic theory for populations that are not infinite and in which mating is not at random. Individuals with common

Box 4.6: Wahlund effect
Suppose subpopulations are each in Hardy-Weinberg equilibrium, but with different proportions for their alleles. In subpopulation i the proportion of allele A_1 is p_{1i} and the proportion of the A_1A_1 homozygote is $P_{11_i} = p_{1i}^2$. If the whole population consists of a proportion m_i of individuals belonging to subpopulation i, then p_1, the proportion of A_1 in the combined population, is

$$p_1 = \sum_i m_i p_{1i}$$

The proportion P_{11} of the A_1A_1 homozygote in the combined population is

$$\begin{aligned} P_{11} &= \sum_i m_i P_{11i} \\ &= \sum_i m_i p_{1i}^2 \\ &= (\sum_i m_i p_{1i})^2 + \sum_i m_i (p_{1i} - p_1)^2 \\ &= p_1^2 + \sum_i m_i (p_{1i} - p_1)^2 \end{aligned}$$

The second term on the right hand side is positive, so it follows that

$$P_{11} > p_1^2$$

For A_1A_2 heterozygotes, a similar argument leads to

$$\begin{aligned} P_{12} &= \sum_i m_i P_{12i} \\ &= \sum_i m_i 2p_{1i}p_{2i} \\ &= 2p_1p_2 + 2\sum_i m_i (p_{1i} - p_1)(p_{2i} - p_2) \end{aligned}$$

but the second term on the right hand side can be positive or negative.

ancestors are said to be RELATED, and their children are INBRED. If no further qualifications are made, then all humans are both inbred and related to everyone else simply because the population is finite. Each of us has two parents and four grandparents. If we all had eight distinct grandparents, 16 distinct great-grandparents, and so on, it would take only a few hundred years back in time before we would have more ancestors among us than there were people living on the planet at that time. Obviously then our parents have some ancestors in common, but conventional definitions of inbreeding refer only to children whose parents are related through people in the past few generations.

Inbreeding in Pedigrees

The genetic consequences of inbreeding follow directly from basic Mendelian principles. An individual receives half of his or her genetic material from each parent, and transmits half of this total to each child. For each gene, an individual receives two alleles, one from each parent, and is generally equally likely to transmit either of these two alleles to a child. The random element in such transmission means that statements about inbreeding are usually expressed as probabilities. Because related people share ancestors, there is a chance that they each receive a copy of the same allele from one ancestor. Figure 4.2 shows the pedigree for a person I whose parents X and Y are half sibs. The parent that X and Y have in common is H. We first assume that H is not inbred and that all three grandparents G, H, and J of I are unrelated to each other. In the figure, we have labeled the alleles transmitted from H to X and Y as h_1 and h_2, respectively. There is a probability of 0.5 that these two alleles are copies of the same allele–i.e., they both descend from just one of the alleles received by H from his or her parents. We write the probability of this event of IDENTITY BY DESCENT, ibd, as

$$\begin{aligned} \Pr(h_1 \textit{ is ibd to } h_2) &= \Pr(h_1 \equiv h_2) \\ &= 0.5 \end{aligned}$$

We will also use ibd for the phrase "identical by descent." Individuals X and Y, in turn, transmit alleles a and b to their child I, and we are interested in the probability of these two alleles being ibd. For this to occur, first h_1 and h_2 must be ibd, and then X must transmit a copy of h_1 (with probability 0.5), and Y a copy of h_2 (with probability 0.5):

$$\Pr(a \equiv b) = \Pr(a \equiv h_1, b \equiv h_2 | h_1 \equiv h_2) \Pr(h_1 \equiv h_2)$$

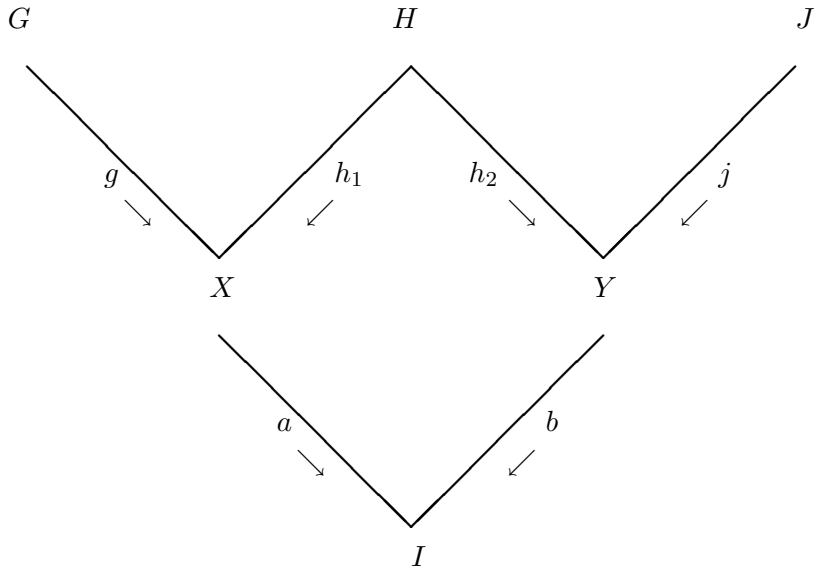


Figure 4.2: I is child of half sibs X and Y

$$\begin{aligned} &= \Pr(a \equiv h_1, b \equiv h_2) \Pr(h_1 \equiv h_2) \\ &= \Pr(a \equiv h_1) \Pr(b \equiv h_2) \Pr(h_1 \equiv h_2) \\ &= 0.5 \times 0.5 \times 0.5 \\ &= 0.125 \end{aligned}$$

Note that the two events of a being a copy of h_1 and b being a copy of h_2 are independent, and they do not depend on the event $h_1 \equiv h_2$. We define the probability that I receives a pair of ibd alleles from his or her parents as the INBREEDING COEFFICIENT F_I of I. So, in this case,

$$F_I = \Pr(a \equiv b) = 0.125$$

Now if H had parents who were related, he or she would have a nonzero inbreeding coefficient F_H, and there would be two ways in which h_1 and h_2 could be ibd:

- With probability 0.5, they descended from the same single allele of the two alleles carried by H.
- With probability 0.5, they descended from the two different alleles in H, and with probability F_H, these alleles were ibd.

Therefore

$$\Pr(h_1 \equiv h_2) \;=\; \frac{1}{2} + \frac{1}{2}F_H$$

so that, following the same argument as before,

$$\begin{aligned}
\Pr(a \equiv b) &= \Pr(a \equiv h_1, b \equiv h_2 | h_1 \equiv h_2) \Pr(h_1 \equiv h_2) \\
&= \Pr(a \equiv h_1, b \equiv h_2) \Pr(h_1 \equiv h_2) \\
&= \Pr(a \equiv h_1) \Pr(b \equiv h_2) \Pr(h_1 \equiv h_2) \\
&= 0.5 \times 0.5 \times (1 + F_H)/2 \\
&= (1 + F_H)/8
\end{aligned}$$

A general approach is to specify some initial or reference population, in which all members are assumed to be unrelated, and then to measure inbreeding relative to that generation. It is generally accepted, for example, that Finland was settled by a relatively small group of people about 4,000 years ago. It would be convenient to quantify inbreeding for the present population as the probability that a random person in the population (assumed to have descended from the initial group) receives two alleles that trace back to a single allele among the founders. Alleles that trace to distinct founding alleles will be considered not ibd since we assumed there was no relatedness among the founders.

The argument given for half sibs leads to PATH-COUNTING equations for inbreeding coefficients. Suppose the parents X and Y of individual I have a common ancestor A. One of these ancestors is shown in Figure 4.3, although there may be several and they need not all be in the same generation. Also suppose that there are n_A people in the loop from one parent through A and back to the other parent. Then, summing over all common ancestors A of X and Y, the inbreeding coefficient of I is

$$F_I \;=\; \sum_A \left(\frac{1}{2}\right)^{n_A} (1 + F_A) \tag{4.6}$$

Each person in the loop, apart from A but including the parents X and Y, introduces another probability of 0.5 for the transmission of an allele from A. The two alleles that common ancestor A gives to the two sides of the loop have a probability $(1 + F_A)/2$ of being ibd.

In the half sib case of Figure 4.2, parent H is the only common ancestor of X and Y and $n_H = 3$ for the path XHY, so $F_I = 1/8$ as before. In Figure 4.4, full sibs X and Y have two parents G and H in common. The two

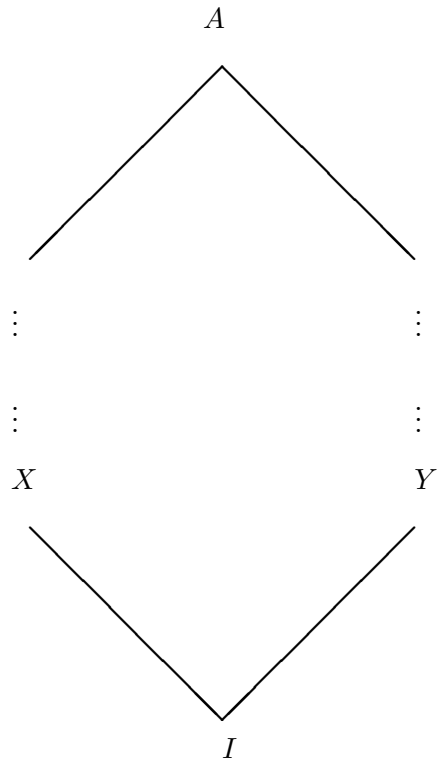


Figure 4.3: The parents X and Y of I have a common ancestor A.

paths are XGY and XHY, each with three individuals, so the inbreeding coefficient of their child I would be $(1/2)^3 + (1/2)^3 = 1/4$, providing G and H are not inbred. In Figure 4.5, first cousins X and Y have four distinct parents, G, H, J and K, two of whom are full sibs, so they have two grandparents A and B in common. The two paths $XHAJY$ and $XHBJY$ each have five individuals. The inbreeding coefficient of the children of first cousins is therefore $(1/2)^5+(1/2)^5 = 1/16$, and this is the maximum amount of inbreeding tolerated by most marriage laws.

Just as the concepts of inbreeding and relatedness are closely connected, so are the probabilities of these events. The usual measure of relatedness for individuals X and Y is the COANCESTRY COEFFICIENT θ_{XY}, defined as the probability that two alleles, one taken at random from each of X and Y, are ibd. If a and b are the alleles from X and Y, then

$$\theta_{XY} = \Pr(a \equiv b | a, b \textit{ have come from } X \textit{ and } Y)$$

If individuals X and Y have a child I,

$$F_I = \theta_{XY} \tag{4.7}$$

so we have shown in the preceding examples that θ_{XY} is 1/4 for full sibs, 1/8 for half sibs, and 1/16 for first cousins.

There is one additional relation needed to characterize inbreeding in pedigrees or in populations. That is the probability of two alleles from the same individual being ibd. For individual X we have already defined F_X as the probability that X *receives* two ibd alleles. Now we introduce the probability θ_{XX} that X *transmits* two ibd alleles, and this is

$$\theta_{XX} = \Pr(a \equiv b | a, b \textit{ are both from } X)$$

In the examples we have done so far, we have assumed that each individual is not inbred, and if X is not inbred then $\theta_{XX} = 0.5$. However, there will be occasions where we need to allow for inbred individuals. Consider individual X in Figure 4.6, who receives alleles c and d from his parents and who transmits alleles a and b to two of his children. There are four possibilities, each of which has probability 1/4:

- a and b are both copies of c.
- a is a copy of c, and b is a copy of d.
- a is a copy of d, and b is a copy of c.
- a and b are both copies of d.

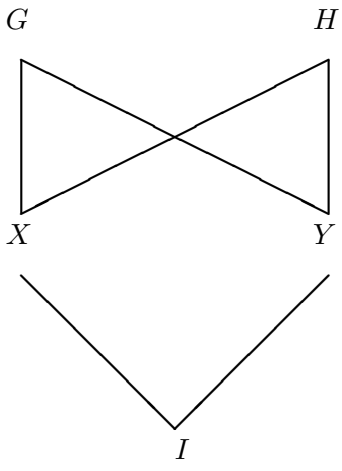


Figure 4.4: I is child of full sibs X and Y.

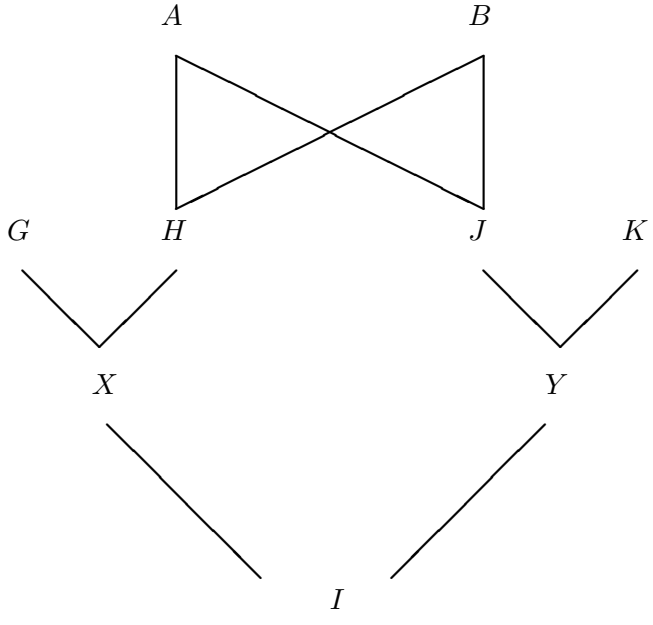


Figure 4.5: I is child of first cousins X and Y.

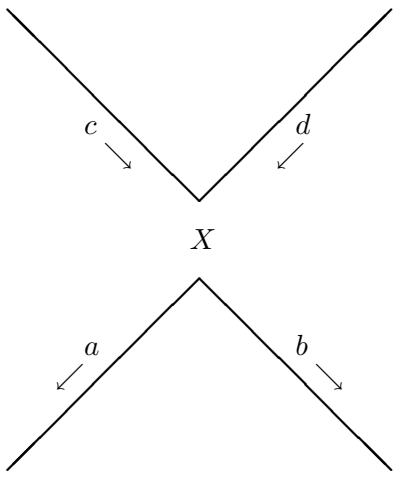


Figure 4.6: $\Pr(a \equiv b)$ is coancestry of an individual with itself.

Using the law of total probability,

$$\begin{aligned} \theta_{XX} &= \Pr(a \equiv b) \\ &= \frac{1}{4}[\Pr(c \equiv c) + \Pr(c \equiv d) + \Pr(d \equiv c) + \Pr(d \equiv d)] \end{aligned}$$

Now $\Pr(c \equiv c) = \Pr(d \equiv d) = 1$ and $\Pr(c \equiv d) = \Pr(d \equiv c)$, which is just the inbreeding coefficient of X. Making these substitutions

$$\theta_{XX} = \frac{1}{2}(1 + F_X) \tag{4.8}$$

We have already met the result in Equation 4.8 in setting up the path-counting rule for inbreeding coefficients. The alleles the individual I received eventually traced back to common ancestor A, and the probability of them then being ibd is $\theta_{AA} = (1 + F_A)/2$.

Inbreeding in Populations

We have defined the inbreeding coefficient and coancestry by referring to specific individuals. However, when we are dealing with groups of people we use average values of F and θ, written without subscripts. We have to

return to the idea of picking people at random from the population. In this sense, F is the probability that $a \equiv b$ given that a and b are the two alleles of a person chosen at random from the population.

$$F = \Pr(a \equiv b | a, b \textit{ are the two alleles of a person chosen at random})$$

Similarly, θ is the probability that $a \equiv b$ given that a and b are two alleles, one from each of two people selected at random from the population.

$$\theta = \Pr(a \equiv b | a, b \textit{ are from two people chosen at random})$$

Sometimes θ is written as F_{ST}.

In a finite population, there is an increasing chance over time that any pair of alleles will be identical by descent, simply because some alleles in each generation will not be passed to the next generation whereas other alleles will have multiple copies passed on. To quantify this increase in inbreeding, it is convenient first to set up equations for MONOECIOUS populations in which selfing is allowed. Any individual can mate with any other, and can even self-mate by providing both copies of each allele to a child. In this system, there is complete random pairing of alleles in each generation. Many plant species are monoecious.

If the population has a level of inbreeding of F, what will be the inbreeding level in the next generation? Because of our assumption of completely random mating in populations of size N, the probability that a member of the next generation has a single parent X is $1/N$, and the probability he or she has distinct parents X and Y is $1 - 1/N$. If the individual has a single parent X, then the probability he or she will receive two ibd alleles is the coancestry of X with itself, and we saw in Equation 4.8 that this is $(1 + F_X)/2$. If the individual has two parents, X and Y, then the probability he or she receives two ibd alleles is their coancestry $\theta_{XY} = \theta$. Recall that F and θ refer to random individuals or random pairs of individuals, respectively, in the population. So, if we now use a and b to denote the alleles received by an individual in the next generation, we can define the new inbreeding coefficient, F', as

$$\begin{aligned}
F' &= \Pr(a \equiv b) \\
&= \Pr(a \equiv b | a, b \textit{ came from the same individual}) \\
&\quad \times \Pr(a, b \textit{ came from the same individual}) \\
&\quad + \Pr(a \equiv b | a, b \textit{ came from different individuals}) \\
&\quad \times \Pr(a, b \textit{ came from different individuals}) \\
&= \frac{1+F}{2} \times \frac{1}{N} + \theta \times \left(1 - \frac{1}{N}\right)
\end{aligned}$$

Because the special conditions we have set include selfing, the descent status of a pair of alleles is the same whether they are in the same or different individuals; that is, under these conditions $F = \theta$. Therefore, we can write $F' = \theta'$ and rearrange the last equation to give

$$\theta' = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right)\theta$$

If there is no inbreeding or relatedness in the generation $t = 0$, and if $\theta_1, \theta_2, \ldots, \theta_t$ denote the inbreeding and coancestry coefficients in subsequent generations, then

$$\theta_1 = \frac{1}{2N}$$

$$\theta_2 = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right)\frac{1}{2N} = 1 - \left(1 - \frac{1}{2N}\right)^2$$

and so on, until

$$\theta_t = 1 - \left(1 - \frac{1}{2N}\right)^t \tag{4.9}$$

In Figure 4.7 we show how $F = \theta$ changes over time for three values of N for this case of random mating. We will see that the increase in F or θ accompanies a decrease in genetic variation within a population, and this is the process of genetic drift referred to earlier. Clearly the rate of increase is very low for large values of N. Note that N is the "effective" size of the population, referring to the size of the group of possible parents of an individual. In this simplified model, the effective number can be regarded as the size of the population from which a person chooses a mate. This will be much smaller than the census size of the population, and the historical value is generally thought to be about 100,000 for human populations.

Exercise 4.6 Simulate the process of genetic drift for five generations for a population of size 5 initiated with 10 different alleles. In each generation, number the alleles from 1 to 10, and note their allelic state. Take 10 successive digits from an arbitrary starting point in a table of random numbers, and regard these as the parental alleles. For example, if the parental generation is

Number	1	2	3	4	5	6	7	8	9	10
Allele	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}

and Appendix Table A.3 of random numbers is entered in column 12 of row 31, the set of random digits is 9309 16023 0 and generation one will have alleles

Number	1	2	3	4	5	6	7	8	9	10
Allele	A_9	A_3	A_{10}	A_9	A_1	A_6	A_{10}	A_2	A_3	A_{10}

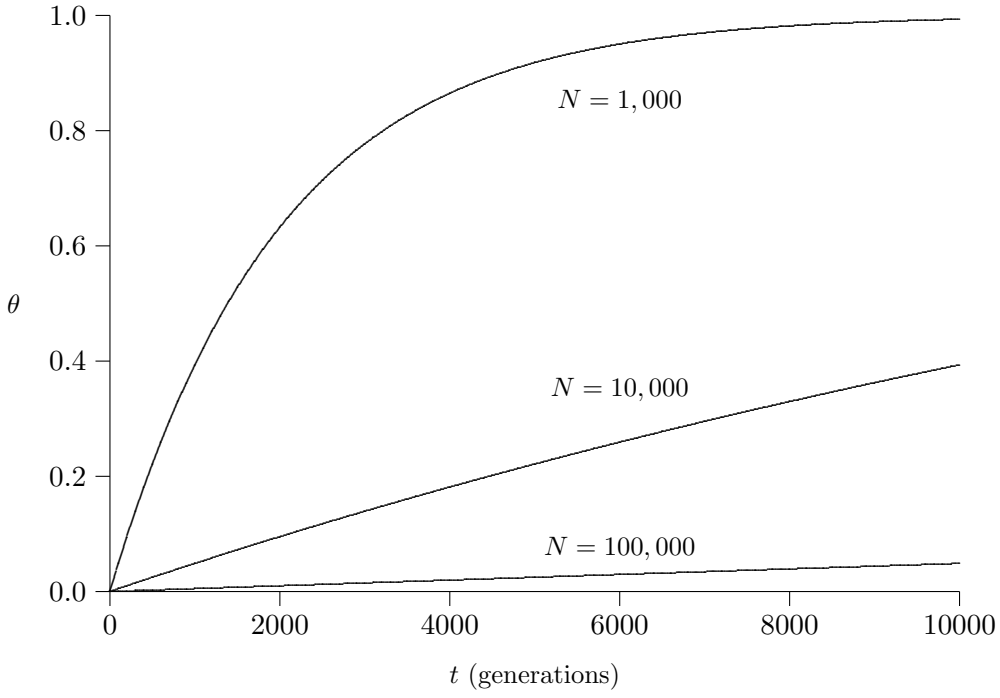


Figure 4.7: Change in θ over time with drift.

Already, four alleles have been lost. The next set of random digits in Table A.3 is 6062 10840 2, so generation two will have alleles

N*umber*	1	2	3	4	5	6	7	8	9	10
A*llele*	A_6	A_{10}	A_6	A_3	A_9	A_{10}	A_2	A_9	A_{10}	A_3

and now there are only five allelic types remaining.

Exercise 4.7 Calculate the actual inbreeding coefficient for the simulations in Exercise 4.6 as the proportion of individuals with ibd alleles in each generation. To do this, take alleles 1 and 2 to represent individual 1, alleles 3 and 4 to represent individual 2, and so on. The theoretical values are 0.00, 0.10, 0.19, 0.27, 0.34, and 0.41 for $t = 0, 1, 2, 3, 4$, and 5.

The calculations in Box 4.7 show that the Equation 4.9 is approximately true even when selfing is not possible. Numerical results for the two sets of expressions are shown in Table 4.3. It can be seen that, for large population sizes, there is no discernible difference in the inbreeding levels for the two mating systems. More elaborate theory can be developed for separate sexes, but the result is the same - population genetic theory for human populations can be approximated by theory for monoecious populations.

Genetic drift, unlike mutation or selection, does not change the overall average, or EXPECTED, allele proportions. The average proportion of an allele over many replicate populations (see Figure 4.1) remains the same, although it will certainly change within any single population. There is therefore a great deal of variation generated between populations. A series of finite isolated populations, each descending from the same reference population, will drift apart over time.

Genotype Proportions in Inbred Populations

Another way of expressing the phenomenon of genetic drift is to express genotypic proportions in terms of allelic proportions *and* the inbreeding coefficient. An individual will be homozygous if it receives two copies of the same allele. With probability F these two are ibd, and therefore are both allele A if either one is A. With probability $(1 - F)$ the two are not ibd, and each has chance p_A of being allele A:

$$P_{AA} = Fp_A + (1 - F)p_A^2$$

Box 4.7: Avoidance of selfing

What relevance does Equation 4.9 have for human populations, where selfing is not possible? As a partial demonstration that it is a very good approximation, consider a population of size N where mating is at random but selfing is not possible. A child in generation $t+1$ must receive alleles from distinct parents in generation t, and these alleles have chance θ of being ibd:

$$F_{t+1} \quad = \quad \theta_t$$

Alleles received by different children, however, can come from the same or different parents with probabilities $1/N$ and $(N-1)/N$, so that

$$\theta_{t+1} \quad = \quad \frac{1}{N}\frac{1+F_t}{2} + \frac{N-1}{N}\theta_t$$

Putting these two equations together provides

$$F_{t+2} \quad = \quad \frac{1}{2N} + \frac{N-1}{N}F_{t+1} + \frac{1}{2N}F_t$$

For the case where individuals in the initial population are noninbred and unrelated, this leads to

$$F_t \quad = \quad 1 - \left[\frac{(1-\lambda_2)\lambda_1^t - (1-\lambda_1)\lambda_2^t}{\lambda_1 - \lambda_2}\right]$$

Here $\lambda_1, \lambda_2 \quad = \quad [(N-1)/N \pm \sqrt{(1+1/N^2)}]/2$ and when N is large, $\lambda_1 \approx (1-1/2N), \lambda_2 \approx 0$. In that case the result reduces to Equation 4.9, derived for random mating when selfing is allowed.

Table 4.3: Inbreeding coefficients for monoecious populations of size $N = 100,000$.

Generation t	*With selfing* $F = \theta$	*Without selfing* F	θ
$1,000$	0.0050	0.0050	0.0050
$10,000$	0.0488	0.0488	0.0488
$100,000$	0.3935	0.3935	0.3935
$1,000,000$	0.9933	0.9933	0.9933
$10,000,000$	1.0000	1.0000	1.0000

or, generally,

$$\left.\begin{array}{rcl} P_{ii} & = & p_i^2 + p_i(1-p_i)F \\ P_{ij} & = & 2p_ip_j(1-F) \end{array}\right\} \tag{4.10}$$

showing that genotypic proportions can be expressed as the Hardy-Weinberg values plus a deviation due to drift. In fact, this deviation can be due to any system of INBREEDING–meaning a mating system in which an individual can receive ibd allele pairs. Note that these equations assume the initial population to be in Hardy-Weinberg equilibrium. For random-mating populations, the ibd status of pairs of alleles is the same whether they are carried by the same or different individuals, and then $F = \theta$. For this reason, Equations 4.10 are sometimes written with θ replacing F.

For alleles that are both harmful and recessive, such as the $\Delta F508$ allele responsible for most cases of cystic fibrosis, inbreeding increases the proportion of people with the harmful trait by virtue of having two copies of the deleterious allele. These two alleles are not masked by a normal allele. The $\Delta F508$ allele in Caucasian populations has a proportion of about $p = 0.05$. Among individuals whose parents are unrelated, the probability of having two copies of the allele, and therefore having cystic fibrosis, is about $p^2 = 0.0025$. Among people whose parents are cousins, however, with probability $(1 - F) = 15/16$ the genotype probability is p^2, and with probability $F = 1/16$ it has the higher value of p. The total probability of the disease among these inbred people is more than doubled, to 0.0055.

Homozygotes have two alleles that have the same chemical composition, and so are IDENTICAL IN STATE. Such alleles may or may not be ibd. Heterozygotes have alleles that are not identical in state, and these alleles cannot be ibd.

Drift and Mutation

A previous section showed that mutation and selection could act in opposite ways to lead to an equilibrium proportion for an allele. A different type of equilibrium can be established between drift and mutation. Genetic variation is lost by drift, as alleles tend to become FIXED, whereas variation can continually be introduced by mutation of the form that every mutant is a new type of allele. Equilibrium now refers to a constant amount of variation–the proportion of any particular allele will be changing. Every allele introduced by mutation will eventually be lost or fixed. Any fixation is temporary, however, as further mutations will introduce other alleles.

A convenient way to characterize such populations is by use of the inbreeding coefficient. The transition equation for changes due to drift

$$F' = \frac{1}{2N} + (1 - \frac{1}{2N})F$$

has to be modified. Alleles can remain ibd only if neither of them mutates, and this happens with probability $(1-\mu)^2$, where μ is the mutation rate. Therefore

$$F' = (1-\mu)^2 \left[\frac{1}{2N} + (1 - \frac{1}{2N})F\right]$$

At equilibrium, the inbreeding coefficient is no longer changing and has the value

$$\begin{aligned} \hat{F} &= \frac{(1-\mu)^2/2N}{1-(1-\mu)^2(1-1/2N)} \\ &\approx \frac{1}{1+4N\mu} \end{aligned}$$

An alternative expression is

$$\hat{H} \approx \frac{4N\mu}{1+4N\mu}$$

where $\hat{H}$ is the proportion of heterozygotes in the equilibrium population.

FOUR-ALLELE DESCENT MEASURES

So far, we have considered the genotypes of individuals or of collections of individuals. There are occasions, such as in the "brother's defense" discussed later, where it is necessary to consider the genotypes of pairs of individuals. These pairs may be relatives, such as brothers, or they may both belong to some specific subpopulation. For pairs of people, it is necessary to introduce more ibd measures. Whereas the genotypic proportion of possibly inbred individuals requires the use of a descent measure defined for pairs of alleles, the genotypic proportions for pairs of relatives require descent measures for three or four alleles being ibd. Suppose individual X has alleles a and b at locus **A** and Y has alleles c and d. There are fifteen possible ibd relations among the four alleles, as shown in Table 4.4, along with their probabilities δ. The fifteen probabilities add to one.

We need to explain the notation. For each δ, the subscript indicates which alleles are ibd, and alleles not in the subscript are neither ibd to

Table 4.4: Descent relations among alleles for two individuals: X with alleles a and b and Y with alleles c and d.

	Probability	
Alleles ibd[1]	General	Full sibs[2]
none	δ_o	1/4
$a \equiv b$	δ_{ab}	0
$c \equiv d$	δ_{cd}	0
$a \equiv c$	δ_{ac}	1/4
$a \equiv d$	δ_{ad}	0
$b \equiv c$	δ_{bc}	0
$b \equiv d$	δ_{bd}	1/4
$a \equiv b \equiv c$	δ_{abc}	0
$a \equiv b \equiv d$	δ_{abd}	0
$a \equiv c \equiv d$	δ_{acd}	0
$b \equiv c \equiv d$	δ_{bcd}	0
$a \equiv b, c \equiv d$	$\delta_{ab.cd}$	0
$a \equiv c, b \equiv d$	$\delta_{ac.bd}$	1/4
$a \equiv d, b \equiv c$	$\delta_{ad.bc}$	0
$a \equiv b \equiv c \equiv d$	δ_{abcd}	0
Total	1	1

[1] *Alleles not specified are not ibd.*
[2] a, c *from mother;* b, d *from father.*

each other nor to the alleles specified. For example, the quantity δ_{ab} is the probability that $a \equiv b$ are ibd *and* that c and d are neither ibd to each other nor to a and b. The equivalence sign $\equiv$ is being used to indicate ibd. The quantity $\delta_{ab.cd}$ is the probability that $a \equiv b, c \equiv d$ but that the two pairs a, b and c, d are not ibd to each other. Finally, δ_{abcd} is the probability that all four alleles are ibd.

As an example consider the case where X and Y are full sibs with parents G and H, as was shown in Figure 4.4. We redraw the pedigree in Figure 4.8 to show the four alleles a, b, c and d. If G and H are not inbred and are unrelated, there are only four possible ibd relationships with nonzero probabilities: $a \equiv c, b \equiv d$; $a \equiv c, b \not\equiv d$; $a \not\equiv c, b \equiv d$; and $a \not\equiv c, b \not\equiv d$. Each of these has probability 1/4, as shown in Table 4.4.

What is the relationship between the four-allele descent measures δ and the two-allele measure θ introduced earlier? Recall that θ_{XY} was defined as the probability of a random allele from X being ibd to a random allele from

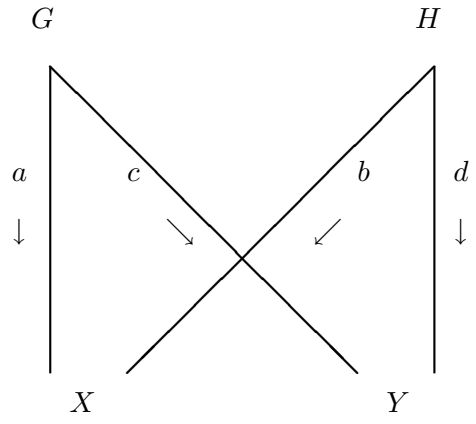


Figure 4.8: Pedigree of full sibs X and Y.

Y. When X has alleles a and b and Y has alleles c and d this means that

$$\theta_{XY} = \frac{1}{4}[\Pr(a \equiv c) + \Pr(a \equiv d) + \Pr(b \equiv c) + \Pr(b \equiv d)]$$

so that

$$\begin{aligned}\theta_{XY} = \frac{1}{4}[&(\delta_{ac} + \delta_{abc} + \delta_{acd} + \delta_{ac.bd} + \delta_{abcd}) \\ &+ (\delta_{ad} + \delta_{abd} + \delta_{acd} + \delta_{ad.bc} + \delta_{abcd}) \\ &+ (\delta_{bc} + \delta_{abc} + \delta_{bcd} + \delta_{bc.ad} + \delta_{abcd}) \\ &+ (\delta_{bd} + \delta_{abd} + \delta_{bcd} + \delta_{bd.ac} + \delta_{abcd})]\end{aligned} \tag{4.11}$$

Noninbred Relatives

Now suppose that neither of two individuals $X(a, b)$ and $Y(c, d)$ is inbred. Then any probability involving $a \equiv b$ or $c \equiv d$ is zero, and there are only seven measures to consider: $\delta_0, \delta_{ac}, \delta_{ad}, \delta_{bc}, \delta_{bd}, \delta_{ac.bd}$, and $\delta_{ad.bc}$. Furthermore, if there is symmetry between a, b and c, d then there are only three distinct values for these seven measures, according to whether the two individuals have zero (δ_0), or one ($\delta_{ac}, \delta_{ad}, \delta_{bc}, \delta_{bd}$), or two ($\delta_{ac.bd}, \delta_{ad.bc}$) alleles identical by descent. There are occasions when it is better to work with all seven δ measures, however.

For noninbred relatives, Equation 4.11 reduces to

$$\theta_{XY} = \frac{1}{4}[(\delta_{ac} + \delta_{ad} + \delta_{bc} + \delta_{bd}) + 2(\delta_{ac.bd} + \delta_{ad.bc})]$$

For full sibs, therefore (as before),

$$\theta_{XY} = \frac{1}{4}[(\frac{1}{4} + 0 + 0 + \frac{1}{4}) + 2(\frac{1}{4} + 0)] = \frac{1}{4}$$

Another useful summary measure in the noninbred case is the two-allele-pair measure $\Delta_{\ddot{X}+\ddot{Y}}$. This is the average of the two probabilities that X and Y have two pairs of ibd alleles:

$$\Delta_{\ddot{X}+\ddot{Y}} = \frac{1}{2}(\delta_{ac.bd} + \delta_{ad.bc})$$

When X and Y are full sibs,

$$\Delta_{\ddot{X}+\ddot{Y}} = \frac{1}{2}(\frac{1}{4} + 0) = \frac{1}{8}$$

The probabilities that noninbred relatives X and Y have 0, 1, or 2 pairs of ibd alleles can be summarized as

$$\left.\begin{array}{rl} 0\ pairs: & 1 - 4\theta_{XY} + 2\Delta_{\ddot{X}+\ddot{Y}} \\ 1\ pair: & 4(\theta_{XY} - \Delta_{\ddot{X}+\ddot{Y}}) \\ 2\ pairs: & 2\Delta_{\ddot{X}+\ddot{Y}} \end{array}\right\} \tag{4.12}$$

Joint Genotypic Probabilities

The descent status of the four alleles that two individuals have between them puts constraints on the possible genotypes of the individuals (Cockerham 1971). If all four alleles were ibd, for example, both individuals would need to be of the same homozygous genotype. The converse relation is more complicated because homozygous individuals need not have alleles that are ibd. We now consider all possible pairs of genotypes for individuals related to any degree (specified by the descent measures in Table 4.4).

Two homozygotes. What is the probability that X and Y are both homozygous A_iA_i? In the third column of Table 4.5 we show contributions to this probability from each of the 15 ibd relationships. With probability δ_0, there is no identity by descent among the four alleles, so each is of independent origin, and each has probability p_i of being of type A_i. In the second row of Table 4.5, only alleles a and b are ibd and have the same origin. This means that there are three alleles with independent origin: ab, c, and d, and each has probability p_i of being of type A_i. The probability of two A_iA_i genotypes in this ibd situation is therefore p_i^3. The overall probability

is arrived at by combining third-column terms over rows of Table 4.5 using the law of total probability, and is

$$\begin{aligned}\Pr(A_iA_i, A_iA_i) &= \delta_{abcd}p_i + (\delta_{abc} + \delta_{abd} + \delta_{acd} + \delta_{bcd})p_i^2 \\ &\quad + (\delta_{ab.cd} + \delta_{ac.bd} + \delta_{ad.bc})p_i^2 \\ &\quad + (\delta_{ab} + \delta_{ac} + \delta_{ad} + \delta_{bc} + \delta_{bd} + \delta_{cd})p_i^3 + \delta_0 p_i^4\end{aligned}$$

The probability with which two individuals are homozygous, but for different alleles, can be derived in a similar manner, using the fourth column of Table 4.5. For any of the rows in which the individuals share ibd alleles, there is zero probability that they could be homozygous for different alleles. Adding the fourth-column terms over rows gives

$$\Pr(A_iA_i, A_jA_j) = \delta_{ab.cd}p_ip_j + \delta_{ab}p_ip_j^2 + \delta_{cd}p_i^2p_j + \delta_0 p_i^2p_j^2$$

When neither X nor Y is inbred, the last two results simplify to

$$\Pr(A_iA_i, A_iA_i) = p_i^4 + 4\theta_{XY}p_i^3(1-p_i) + 2\Delta_{\ddot{X}+\ddot{Y}}p_i^2(1-p_i)^2 \quad (4.13)$$

$$\Pr(A_iA_i, A_jA_j) = (1 - 4\theta_{XY} + 2\Delta_{\ddot{X}+\ddot{Y}})p_i^2p_j^2 \quad (4.14)$$

Under the assumption of no inbreeding, it is quicker to derive these last two results directly from Equations 4.12.

One homozygote and one heterozygote. For one homozygous and one heterozygous individual, we need to distinguish between the cases when the individuals share an allele and when they do not. We also need to allow for both orderings of alleles within heterozygotes, and the calculations follow from Table 4.6.

Adding columns 2 and 3 in Table 4.6, multiplying by column 1, and then summing over rows, we get

$$\begin{aligned}\Pr(A_iA_i, A_iA_j) &= 2\delta_0 p_i^3p_j + (2\delta_{ab} + \delta_{ac} + \delta_{ad} + \delta_{bc} + \delta_{bd})p_i^2p_j \\ &\quad + (\delta_{abc} + \delta_{abd})p_ip_j\end{aligned}$$

and adding over rows for column 1 multiplied by column 4 (doubled to take account of both heterozygote orders A_jA_k, A_kA_j), we get

$$\Pr(A_iA_i, A_jA_k) = 2\delta_0 p_i^2p_jp_k + \delta_{ab}p_ip_jp_k$$

Table 4.5: Probabilities of $X(a,b)$ and $Y(c,d)$ being the same homozygote A_iA_i or different homozygotes A_iA_i, A_jA_j.

$\mathrm{I}BD$[1]	$\Pr(\mathrm{I}BD)$	$\Pr(A_iA_i, A_iA_i \mid \mathrm{I}BD)$	$\Pr(A_iA_i, A_jA_j \mid \mathrm{I}BD))$
none	δ_o	p_i^4	$p_i^2 p_j^2$
$a \equiv b$	δ_{ab}	p_i^3	$p_i p_j^2$
$c \equiv d$	δ_{cd}	p_i^3	$p_i^2 p_j$
$a \equiv c$	δ_{ac}	p_i^3	0
$a \equiv d$	δ_{ad}	p_i^3	0
$b \equiv c$	δ_{bc}	p_i^3	0
$b \equiv d$	δ_{bd}	p_i^3	0
$a \equiv b \equiv c$	δ_{abc}	p_i^2	0
$a \equiv b \equiv d$	δ_{abd}	p_i^2	0
$a \equiv c \equiv d$	δ_{acd}	p_i^2	0
$b \equiv c \equiv d$	δ_{bcd}	p_i^2	0
$a \equiv b, c \equiv d$	$\delta_{ab.cd}$	p_i^2	$p_i p_j$
$a \equiv c, b \equiv d$	$\delta_{ac.bd}$	p_i^2	0
$a \equiv d, b \equiv c$	$\delta_{ad.bc}$	p_i^2	0
$a \equiv b \equiv c \equiv d$	δ_{abcd}	p_i	0

[1] $\mathrm{I}BD$: i*bd relationship*

Table 4.6: Probabilities of $X(a,b)$ being homozygous A_iA_i and $Y(c,d)$ being heterozygous A_iA_j or A_jA_k.

	$\Pr(A_iA_i, A_iA_j \mid \mathrm{I}BD)$		
$\Pr(\mathrm{I}BD)$	$c = A_i, d = A_j$	$c = A_j, d = A_i$	$\Pr(A_iA_i, A_jA_k \mid \mathrm{I}BD))$
δ_0	$p_i^3 p_j$	$p_i^3 p_j$	$p_i^2 p_j p_k$
δ_{ab}	$p_i^2 p_j$	$p_i^2 p_j$	$p_i p_j p_k$
δ_{cd}	0	0	0
δ_{ac}	$p_i^2 p_j$	0	0
δ_{ad}	0	$p_i^2 p_j$	0
δ_{bc}	$p_i^2 p_j$	0	0
δ_{bd}	0	$p_i^2 p_j$	0
δ_{abc}	$p_i p_j$	0	0
δ_{abd}	0	$p_i p_j$	0
δ_{acd}	0	0	0
δ_{bcd}	0	0	0
$\delta_{ab.cd}$	0	0	0
$\delta_{ac.bd}$	0	0	0
$\delta_{ad.bc}$	0	0	0
δ_{abcd}	0	0	0

If X is not inbred, and since heterozygote Y cannot be inbred, these two equations become

$$\begin{aligned}\Pr(A_iA_i, A_iA_j) &= 2(1 - 4\theta_{XY} + 2\Delta_{\ddot{X}+\ddot{Y}})p_i^3 p_j \\ &\quad + 4(\theta_{XY} - \Delta_{\ddot{X}+\ddot{Y}})p_i^2 p_j \qquad (4.15)\end{aligned}$$

$$\Pr(A_iA_i, A_jA_k) = 2(1 - 4\theta_{XY} + 2\Delta_{\ddot{X}+\ddot{Y}})p_i^2 p_j p_k \qquad (4.16)$$

and these results also follow directly from Equations 4.12.

The same two heterozygotes. When two individuals have the same heterozygous genotype, our approach requires an accounting for all four orders of the two alleles within each. Combining the last four columns of Table 4.7, multiplying by column 1, and summing over rows leads to

$$\begin{aligned}\Pr(A_iA_j, A_iA_j) &= 4\delta_0 p_i^2 p_j^2 \\ &\quad + (\delta_{ac} + \delta_{ad} + \delta_{bc} + \delta_{bd})p_i p_j (p_i + p_j) \\ &\quad + 2(\delta_{ac.bd} + \delta_{ad.bc})p_i p_j\end{aligned}$$

This can be rewritten in a way that reflects the fact that heterozygotes cannot be inbred:

$$\begin{aligned}\Pr(A_iA_j, A_iA_j) &= 4(1 - 4\theta_{XY} + 2\Delta_{\ddot{X}+\ddot{Y}})p_i^2 p_j^2 \\ &\quad + 4(\theta_{XY} - \Delta_{\ddot{X}+\ddot{Y}})p_i p_j (p_i + p_j) \\ &\quad + 4\Delta_{\ddot{X}+\ddot{Y}} p_i p_j \qquad (4.17)\end{aligned}$$

This equation could also be derived directly from Equations 4.12.

Two different heterozygotes. Finally we consider the case of two heterozygotes that have one or no alleles in common. With one allele in common,

$$\begin{aligned}\Pr(A_iA_j, A_iA_k) &= 4\delta_0 p_i^2 p_j p_k \\ &\quad + (\delta_{ac} + \delta_{ad} + \delta_{bc} + \delta_{bd})p_i p_j p_k\end{aligned}$$

which, after recognizing that both individuals must be noninbred, gives

$$\begin{aligned}\Pr(A_iA_j, A_iA_k) &= 4(1 - 4\theta_{XY} + 2\Delta_{\ddot{X}+\ddot{Y}})p_i^2 p_j p_k \\ &\quad + 4(\theta_{XY} - \Delta_{\ddot{X}+\ddot{Y}})p_i p_j p_k \qquad (4.18)\end{aligned}$$

Table 4.7: Probabilities of $X(a,b)$ and $Y(c,d)$ being heterozygous A_iA_j.

	$\Pr(A_iA_j, A_iA_j \mid \mathrm{I}BD)$			
	$a=A_i, b=A_j$	$a=A_i, b=A_j$	$a=A_j, b=A_i$	$a=A_j, b=A_i$
$\Pr(\mathrm{I}BD)$	$c=A_i, d=A_j$	$c=A_j, d=A_i$	$c=A_i, d=A_j$	$c=A_j, d=A_i$
δ_0	$p_i^2p_j^2$	$p_i^2p_j^2$	$p_i^2p_j^2$	$p_i^2p_j^2$
δ_{ab}	0	0	0	0
δ_{cd}	0	0	0	0
δ_{ac}	$p_ip_j^2$	0	0	$p_i^2p_j$
δ_{ad}	0	$p_ip_j^2$	$p_i^2p_j$	0
δ_{bc}	0	$p_i^2p_j$	$p_ip_j^2$	0
δ_{bd}	$p_i^2p_j$	0	0	$p_ip_j^2$
δ_{abc}	0	0	0	0
δ_{abd}	0	0	0	0
δ_{acd}	0	0	0	0
δ_{bcd}	0	0	0	0
$\delta_{ab.cd}$	0	0	0	0
$\delta_{ac.bd}$	p_ip_j	0	0	p_ip_j
$\delta_{ad.bc}$	0	p_ip_j	p_ip_j	0
δ_{abcd}	0	0	0	0

When the two heterozygotes share no alleles, only the complete noniden-tity probability δ_0 is needed. Multiplying by four to account for all four orders of alleles within individuals gives us

$$\Pr(A_iA_j, A_kA_l) \quad = \quad 4\delta_0 p_i p_j p_k p_l$$

and, because the individuals are both noninbred,

$$\Pr(A_iA_j, A_kA_l) \quad = \quad 4(1 - 4\theta_{XY} + 2\Delta_{\ddot{X}+\ddot{Y}})p_i p_j p_k p_l$$

as may also be seen from Equations 4.12.

Genotypes of Two Sibs

As an application of these results, consider the possible genotype pairs for two sibs X and Y that have noninbred and unrelated parents, so they themselves are not inbred. The descent measure values are shown in Table 4.4, and the probabilities for any of the six possible pairs of genotypes (regardless of order) can be found from Equations 4.13 to 4.18:

A_iA_i, A_iA_i	$p_i^2(1+p_i)^2/4$
A_iA_i, A_jA_j	$p_i^2p_j^2/2$
A_iA_i, A_iA_j	$p_i^2p_j(1+p_i)$
A_iA_i, A_jA_k	$2p_i^2p_jp_k$
A_iA_j, A_iA_j	$p_ip_j(1+p_i+p_j+2p_ip_j)/2$
A_iA_j, A_iA_k	$p_ip_jp_k(1+2p_i)$
A_iA_j, A_kA_l	$2p_ip_jp_kp_l$

These are not the same as simply multiplying together the proportions of each of the two genotypes. Relatives are more likely than random members of the population to have the same genotype. Adding together the first and fifth of these equations and then summing over all alleles A_i, A_j gives the probability that two (untyped) sibs have the same genotype without specifying that genotype. This probability is

$$\Pr(sibs\ the\ same) \quad = \quad \frac{1}{4}\left(1 + 2\sum_i p_i^2 + 2(\sum_i p_i^2)^2 - \sum_i p_i^4\right)$$

This expression tends to 1/4 as the number of alleles increases and the proportion of each one becomes small.

Exercise 4.8 Find the probability with which a noninbred uncle and his noninbred nephew both have A_1A_2 genotypes if they belong to a population in Hardy-Weinberg equilibrium with proportions p_1 and p_2 for alleles A_1 and A_2.

MATCH PROBABILITIES

We now return to an assumption that we made when considering transfer evidence in Chapter 2. Recall that we arrived at Equation 2.3 as the expression for the LR in a single crime scene stain case:

$$LR = \frac{1}{\Pr(G_C|G_S, H_d, I)}$$

We assumed that knowledge of the suspect's genotype G_S did not influence our uncertainty about the offender's genotype. Cases arise, however, where that assumption is not correct. It is not correct, for example, when it is suggested that the offender is a close relative of the suspect. It is also not correct when the suspect and offender should realistically be considered both members of the same subpopulation within the population from which the database for estimating allele proportions has been constructed. In these cases, knowledge of the suspect's genotype *does* influence our uncertainty about that of the offender and we now explain how to take account of this knowledge.

To maintain consistency with notation of the previous section, we refer to the offender as X and the suspect as Y. Consider first the case where the suspect has been found to have heterozygous genotype A_iA_j and matches the genotype of the stain presumed to have come from the offender. We require $\Pr(G_X = A_iA_j|G_Y = A_iA_j)$, which in Chapter 2 we wrote as $\Pr(G_C = A_iA_j|G_S = A_iA_j, H_d, I)$. From the third law of probability, we note that

$$\Pr(G_X = A_iA_j|G_Y = A_iA_j) = \frac{\Pr(G_X = A_iA_j, G_Y = A_iA_j)}{\Pr(G_Y = A_iA_j)} \quad (4.19)$$

The joint probability in the numerator of the right hand side of this equation was found in the previous section.

Full Sibs

If X and Y are full sibs, we have found that

$$\begin{aligned}\Pr(G_X = A_iA_i, G_Y = A_iA_i) &= p_i^2(1+p_i)^2/4\\ \Pr(G_X = A_iA_j, G_Y = A_iA_j) &= p_ip_j(1+p_i+p_j+2p_ip_j)/2\end{aligned}$$

The unconditional probabilities that G_Y is A_iA_i or A_iA_j are p_i^2 or $2p_ip_j$, so from Equation 4.19

$$\begin{aligned}\Pr(G_X = A_iA_i|G_Y = A_iA_i) &= (1+p_i)^2/4\\ \Pr(G_X = A_iA_j|G_Y = A_iA_j) &= (1+p_i+p_j+2p_ip_j)/4\end{aligned}$$

Other Relatives

To evaluate joint genotype probabilities for relatives other than sibs, we show the four-allele descent measures in Table 4.8 for many common relationships. Determining the match probabilities is left as an exercise.

Exercise 4.9 Determine the probability with which one relative has a specific homozygous or heterozygous genotype given that the other relative has this type, for each of the following pairs of relatives: (a) Parent-child; (b) Grandparent-grandchild; (c) Half sibs; (d) Uncle-nephew; (e) First cousins.

Unrelated Members of the Same Subpopulation

In the previous sections we have used δ terms that were specific for a precisely defined relationship between two individuals. We now consider the case where X and Y are unrelated individuals that have been selected at random from a subpopulation. In this situation, we can work with average values for the δ terms provided mating is at random in the subpopulation. This means that the ibd relationships among a set of alleles do not depend on how the alleles are arranged within individuals: δ_{ab} has the same value whether a and b are in the same or different individuals, for example. For alleles at the same locus, the four average measures are:

- θ is the probability that any two alleles a and b, selected at random from a subpopulation, are ibd: $\theta = \Pr(a \equiv b)$.
- γ is the probability that any three alleles a, b, and c, selected at random from a subpopulation, are ibd: $\gamma = \Pr(a \equiv b \equiv c)$.
- δ is the probability that any four alleles a, b, c, and d, selected at random from a subpopulation, are ibd: $\delta = \Pr(a \equiv b \equiv c \equiv d)$.
- Δ is the probability that any two pairs of alleles a, b and c, d, selected at random from a subpopulation, are ibd: $\Delta = \Pr(a \equiv b \; and \; c \equiv d)$.

It can be shown (Weir 1994) that these 4 summary measures allow the 15 descent measures for 4 alleles a, b, c and d in Table 4.4 to be evaluated from

$$\begin{aligned} \delta_0 &= 1 - 6\theta + 8\gamma + 3\Delta - 6\delta \\ \delta_{ab} = \delta_{ac} = \delta_{ad} = \delta_{bc} = \delta_{bd} = \delta_{cd} &= \theta - 2\gamma - \Delta + 2\delta \\ \delta_{abc} = \delta_{abd} = \delta_{acd} = \delta_{bcd} &= \gamma - \delta \\ \delta_{ab.cd} = \delta_{ac.bd} = \delta_{ad.bc} &= \Delta - \delta \\ \delta_{abcd} &= \delta \end{aligned}$$

Table 4.8: Descent measures for common relationships.

Probability	Parent-child[1]	Grandparent--grandchild[2]	Full sibs[3]	Half sibs[4]	Uncle-nephew[5]	First cousins[6]
δ_0	0	1/2	1/4	1/2	1/2	3/4
δ_{ab}	0	0	0	0	0	0
δ_{cd}	0	0	0	0	0	0
δ_{ac}	1/2	1/4	1/4	0	1/4	0
δ_{ad}	0	0	0	0	0	0
δ_{bc}	1/2	1/4	0	1/2	1/4	1/4
δ_{bd}	0	0	1/4	0	0	0
δ_{abc}	0	0	0	0	0	0
δ_{abd}	0	0	0	0	0	0
δ_{acd}	0	0	0	0	0	0
δ_{bcd}	0	0	0	0	0	0
$\delta_{ab.cd}$	0	0	0	0	0	0
$\delta_{ac.bd}$	0	0	1/4	0	0	0
$\delta_{ad.bc}$	0	0	0	0	0	0
δ_{abcd}	0	0	0	0	0	0
θ_{XY}	1/4	1/8	1/4	1/8	1/8	1/16
$\Delta_{\ddot{X}+\ddot{Y}}$	0	0	1/8	0	0	0

[1] child Y receives c from parent X.
[2] grandchild Y receives c from grandparent X.
[4] X and Y receive a, c from one parent and b, d from the other.
[4] common parent transmits b to X and c to Y.
[5] nephew Y receives b from full sib of uncle X.
[6] X receives b and Y receives c from full sibs.

Substituting these values of the 15 descent measures into Tables 4.5 and 4.7 provides the following joint probabilities for random individuals X and Y in the same subpopulation:

$$\begin{aligned}\Pr(G_X = A_iA_i, G_Y = A_iA_i) &= (1 - 6\theta + 8\gamma + 3\Delta - 6\delta)p_i^4 \\ &\quad + 6(\theta - 2\gamma - \Delta + 2\delta)p_i^3 \\ &\quad + (4\gamma + 3\Delta - 7\delta)p_i^2 + \delta p_i\end{aligned}$$

$$\begin{aligned}\Pr(G_X = A_iA_j, G_Y = A_iA_j) &= 4(1 - 6\theta + 8\gamma + 3\Delta - 6\delta)p_i^2p_j^2 \\ &\quad + 4(\theta - 2\gamma - \Delta + 2\delta)p_ip_j(p_i + p_j) \\ &\quad + 4(\Delta - \delta)p_ip_j, \;\; i \neq j\end{aligned}$$

The single-individual genotypic proportions for a random-mating population are

$$\begin{aligned}\Pr(G_Y = A_iA_i) &= p_i^2 + p_i(1 - p_i)\theta \\ \Pr(G_Y = A_iA_j) &= 2p_ip_j(1 - \theta)\end{aligned}$$

so that the probabilities of the genotype of X conditional on the genotype of Y are

$$\begin{aligned}\Pr(G_X = A_iA_i|G_Y = A_iA_i) &= [(1 - 6\theta + 8\gamma + 3\Delta - 6\delta)p_i^3 \\ &\quad + 6(\theta - 2\gamma - \Delta + 2\delta)p_i^2 \\ &\quad + (4\gamma + 3\Delta - 7\delta)p_i + \delta]/[p_i + (1 - p_i)\theta]\end{aligned}$$

$$\begin{aligned}\Pr(G_X = A_iA_j|G_Y = A_iA_j) &= 2[(1 - 6\theta + 8\gamma + 3\Delta - 6\delta)p_ip_j \\ &\quad + (\theta - 2\gamma - \Delta + 2\delta)(p_i + p_j) \\ &\quad + (\Delta - \delta)]/(1 - \theta), \;\; i \neq j\end{aligned}$$

For populations in an evolutionary equilibrium, meaning that the quantities θ, γ, δ, and Δ are not changing over time, Li (1996) showed that

$$\begin{aligned}\gamma &= \frac{2\theta^2}{1 + \theta} \\ \delta &= \frac{6\theta^3}{(1 + \theta)(1 + 2\theta)} \\ \Delta &= \frac{\theta^2(1 + 5\theta)}{(1 + \theta)(1 + 2\theta)}\end{aligned}$$

and substituting these into the above conditional probabilities leads to the results of Balding and Nichols (1994):

$$\begin{aligned} \Pr(G_X = A_iA_i|G_Y = A_iA_i) &= \frac{[2\theta + (1-\theta)p_i][3\theta + (1-\theta)p_i]}{(1+\theta)(1+2\theta)} \\ \Pr(G_X = A_iA_j|G_Y = A_iA_j) &= \frac{2[\theta + (1-\theta)p_i][\theta + (1-\theta)p_i]}{(1+\theta)(1+2\theta)} \end{aligned} \tag{4.20}$$

It is Equations 4.20 that should be used in the general case when the assumption $\Pr(G_C|G_S, H_d, I) = \Pr(G_C|H_d, I)$ is dubious. The equation forms the basis of one of the recommendations of the 1996 report of the United States National Research Council. It is meant to be used when two people, the suspect and the person who left the crime sample, both belong to the same subpopulation but allele proportions are not available for that subpopulation. The quantity θ describes the degree of relationship of pairs of alleles within the subpopulation relative to the total population. Allele independence, e.g., Hardy-Weinberg equilibrium, is assumed to hold within subpopulations, but differences in allele proportions among subpopulations mean that there are departures from independence in the whole population. In other words, this treatment explicitly allows for inbreeding and relatedness for all individuals in the population, so that Hardy-Weinberg equilibrium is assumed not to hold at the population level. However, inbreeding levels and relatedness levels for individuals in different families will generally be low.

Effects of θ Corrections

What effect does the use of Equations 4.20 instead of the simple product rule have on the interpretation of matching DNA profiles? The largest effect is the conceptual one of including information on the suspect's genotype in determining the probability of the offender having that type. The numerical effects are small unless allele proportions are small and θ is large.

For heterozygotes A_iA_j in a population with allele proportions $p_i = p_j = p$ and with structure characterized by θ, the likelihood ratio is (from Equations 4.20)

$$\mathrm{L}R = \frac{(1+\theta)(1+2\theta)}{2[\theta + (1-\theta)p]^2}$$

Some numerical values of this expression are

	$\theta = 0$	$\theta = 0.001$	$\theta = 0.01$	$\theta = 0.03$
$p = 0.01$	5,000	4,152	1,295	346
$p = 0.05$	200	193	145	89
$p = 0.10$	50	49	43	34

For homozygotes A_iA_i, with $p_i = p$, the effects tend to be slightly greater. The likelihood ratio is

$$\mathrm{L}R \quad = \quad \frac{(1+\theta)(1+2\theta)}{[2\theta+(1-\theta)p][3\theta+(1-\theta)p]}$$

and some numerical values are

	$\theta = 0$	$\theta = 0.001$	$\theta = 0.01$	$\theta = 0.03$
$p = 0.01$	10,000	6,439	863	157
$p = 0.05$	400	364	186	73
$p = 0.10$	100	96	67	37

The NRC recommendation (National Research Council 1996) of using $\theta = 0.03$ in some cases therefore halves the likelihood ratio for DNA profiles *per locus* when allele proportions are 0.10.

The Relative Nature of Coancestries

It appears that Equations 4.20 offer a solution to the problem of allowing for the effects of population structure on matching probabilities. It is necessary, however, to note that the parameter θ is a relative measure and that the p_i terms are allele proportions in the reference population. To explain this remark, it is helpful to return to the expressions for genotype proportions at a locus.

Central to our evolutionary model is the assumption of random mating within subpopulations. If allele A_i has proportion p_{i_S} in a subpopulation, then, for example, A_iA_i homozygotes have proportion $p_{i_S}^2$ in that subpopulation and this is not altered by the fact of having seen the genotype once already in the subpopulation. However, we generally do not have information about the values of the p_{i_S}. If we take expectations over all replicates of the evolutionary process, we have an expected homozygote proportion of

$$\mathcal{E}(p_{i_S}^2) \quad = \quad p_i^2 + p_i(1-p_i)\theta_S \tag{4.21}$$

as given by Cockerham (1969). In this expression, p_i is the average of the allele proportion over all replicates of the evolutionary process (the separate

columns in Figure 4.1) and could be called the reference population proportion. The quantity θ_S measures the relationship of two alleles within the subpopulation assuming that there is no relationship for alleles in different subpopulations. We expect this quantity to increase, or at least remain constant, over time.

We can allow for dependence among subpopulations of the kind that would occur for a set of (say) Caucasian subpopulations that have a single ancestral population P. Alleles from two such subpopulation are more likely to be identical than would be two alleles from subpopulations from different racial groups. In general, we would expect that subpopulations that have had a greater time of evolutionary separation would have a lesser amount of relationship. If P is the population from which a current collection of subpopulations descended, and if the coancestry in that population was θ_P, then we can modify Equation 4.21 to be conditional on the allele proportions p_{i_P} at that time:

$$\mathcal{E}(p_{i_S}^2|p_{i_P}) \quad = \quad p_{i_P}^2 + p_{i_P}(1-p_{i_P})\beta_S \tag{4.22}$$

(Cockerham 1969) where $\beta = (\theta_S - \theta_P)/(1-\theta_P)$. Taking expectations over values of p_{i_P}, i.e. over replicates of the evolutionary process up until the time of subpopulation divergence, leads us from Equation 4.22 back to Equation 4.21 provided we assume zero coancestry in the initial reference population.

The situation we face in forensic science is of having a sample from a population that is actually an amalgamation of several subpopulations. Because the individual p_{i_S} values are not available, we base our theoretical predictions on expected values such as that in Equations 4.21 or 4.22. If β_S was the same for every subpopulation in the population, then the quantity $p_{i_P}(1-p_{i_P})\beta_S$ would be the variance of allele proportions among the subpopulations from the same population. We will see in Chapter 5 that β_S can be estimated from data from the subpopulations, and this quantity is termed the coancestry of a subpopulation relative to the population. To estimate the quantity θ_S by itself, however, we would need data from subpopulations that had been separated ever since the initial reference population (Cockerham and Weir 1987).

Similar remarks apply to Equations 4.20. If θ is the coancestry of the subpopulation of interest, relative to subpopulations that have been separated so long that they are independent, then reference proportion p_i is the average allele proportion over all subpopulations. The power of Equations 4.20 is that they offer a theoretical means of describing the effects of variation among subpopulations.

ARBITRARY SETS OF ALLELES

In the previous sections we have considered genotype probabilities under the assumption of independence among alleles at a locus, and then the effects of departures from independence caused by inbreeding. We also considered joint genotypic probabilities when pairs of individuals are related or are both members of the same population. We now extend that treatment to allow the probability of any set of alleles at a locus to be determined. This will provide an alternative derivation of the conditional probabilities in Equations 4.20, and will allow extensions to more than two people as may be needed for parentage cases (Chapter 6) or the interpretation of mixed stains (Chapter 7).

We need to return to the discussion at the beginning of this chapter. We made the distinction there between genetic and statistical sampling. Statistical sampling refers to the variation among repeated samples from the same population. If a population has allele A_i in proportion p_i, then the probabilities with which random samples of size $2n$ alleles have sample proportions $\tilde{p}_i$ are given by the multinomial distribution (Chapter 3) and this forms the basis of much of the work in Chapter 5. Genetic sampling refers to the variation among replicates of the evolutionary process, and there is no analog to the binomial that holds for all evolutionary scenarios.

When the population can be regarded as having reached an evolutionary equilibrium, then allele proportions follow the Dirichlet distribution (Wright 1951). This means that the probability of a population having alleles A_i with proportions q_i is given by

$$\Pr(\{q_i\}) \quad = \quad \frac{\Gamma(\gamma_.)}{\prod_i \Gamma(\gamma_i)} \prod_i q_i^{(\gamma_i - 1)}$$

where the parameters γ_i are given by $\gamma_i = (1-\theta)p_i/\theta$ and $\gamma_.$ is their sum $\gamma_. = \sum_i \gamma_i = (1-\theta)/\theta$. Here reference proportion p_i is the allele proportion averaged over all replicate populations, and θ is the same quantity that has been discussed in the previous section. It is assumed to be nonzero. The gamma function $\Gamma(x)$ generally has to be evaluated numerically, but has the property that $\Gamma(x+1) = x\Gamma(x)$. If x is an integer, $\Gamma(x) = (x-1)!$.

A property of the Dirichlet distribution is that the probability of a set of alleles, in which type A_i occurs t_i times and $\sum_i t_i = t_.$ is

$$\Pr(\prod_i A_i^{t_i}) \quad = \quad \frac{\Gamma(\gamma_.)}{\Gamma(\gamma_. + t_.)} \prod_i \frac{\Gamma(\gamma_i + t_i)}{\Gamma(\gamma_i)} \tag{4.23}$$

and this result holds no matter how the alleles are distributed among genotypes. For a single allele, the equation reduces to

$$\Pr(A_i) = \frac{\gamma_i}{\gamma_.} = p_i$$

as it should. For two copies of A_i,

$$\begin{aligned} \Pr(A_i^2) &= \frac{\gamma_i(\gamma_i+1)}{\gamma_.(\gamma_.+1)} \\ &= p_i[(1-\theta)p_i+\theta] \end{aligned} \tag{4.24}$$

and for one copy of each of A_i and A_j,

$$\begin{aligned} \Pr(A_iA_j) &= \frac{\gamma_i\gamma_j}{\gamma_.(\gamma_.+1)} \\ &= (1-\theta)p_ip_j \end{aligned} \tag{4.25}$$

The probability of a heterozygote A_iA_j is twice the probability of the allele set A_i, A_j because of the two allele orders A_iA_j and A_jA_i.

For sets of four alleles, if they are all of type A_i,

$$\begin{aligned} \Pr(A_i^4) &= \frac{\gamma_i(\gamma_i+1)(\gamma_i+2)(\gamma_i+3)}{\gamma_.(\gamma_.+1)(\gamma_.+2)(\gamma_.+3)} \\ &= \frac{p_i[(1-\theta)p_i+\theta][(1-\theta)p_i+2\theta][(1-\theta)p_i+3\theta]}{(1+\theta)(1+2\theta)} \end{aligned} \tag{4.26}$$

and this can be regarded as the probability with which two individuals in the population are both homozygous A_iA_i. If two of the alleles are A_i and two are A_j,

$$\begin{aligned} \Pr(A_i^2A_j^2) &= \frac{\gamma_i(\gamma_i+1)\gamma_j(\gamma_j+1)}{\gamma_.(\gamma_.+1)(\gamma_.+2)(\gamma_.+3)} \\ &= \frac{p_i[(1-\theta)p_i+\theta]p_j[(1-\theta)p_j+\theta]}{(1+\theta)(1+2\theta)} \end{aligned} \tag{4.27}$$

and this is one-fourth the probability with which two individuals, X and Y, are heterozygous A_iA_j (allowing for the two possible orders of alleles per individual). Combining Equations 4.24 and 4.26 gives

$$\begin{aligned} \Pr(G_X = A_iA_i | G_Y = A_iA_i) &= \frac{\Pr(A_i^4)}{\Pr(A_i^2)} \\ &= \frac{[(1-\theta)p_i+2\theta][(1-\theta)p_i+3\theta]}{(1+\theta)(1+2\theta)} \end{aligned}$$

as has been given in Equation 4.20. The conditional probability for heterozygotes $\Pr(G_X = A_iA_j|G_Y = A_iA_j)$, also given in Equation 4.20, follows from dividing Equation 4.27 by Equation 4.25.

The full power of the Dirichlet approach becomes apparent for sets of more than four alleles, for which there is no exact descent measure formulation. To anticipate a result needed in Chapter 6, the probability of a woman and an alleged father, if he is not the child's father, both being homozygous A_iA_i and the woman's child receiving paternal allele A_i is

$$\begin{aligned}
\Pr(A_iA_i, A_iA_i, A_i) &= \Pr(A_i^5) \\
&= \frac{\gamma_i(\gamma_i+1)(\gamma_i+2)(\gamma_i+3)(\gamma_i+4)}{\gamma_\cdot(\gamma_\cdot+1)(\gamma_\cdot+2)(\gamma_\cdot+3)(\gamma_\cdot+4)} \\
&= \frac{p_i[(1-\theta)p_i+\theta][(1-\theta)p_i+2\theta]}{(1-\theta)(1+\theta)} \\
&\quad \times \frac{[(1-\theta)p_i+3\theta][(1-\theta)p_i+4\theta]}{(1+2\theta)(1+3\theta)}
\end{aligned}$$

A result of the type needed in Chapter 7 is for the interpretation of a mixed stain having alleles A_i, A_j, and A_k from a victim and her attacker. If the victim is heterozygous A_iA_j and a man suspected of being the attacker is heterozygous A_iA_k, a probability needed if the actual attacker is homozygous A_kA_k is

$$\begin{aligned}
\Pr(A_iA_j, A_iA_k, A_kA_k) &= \Pr(A_i^2A_jA_k^3) \\
&= \frac{\gamma_i(\gamma_i+1)\gamma_j\gamma_k(\gamma_k+1)(\gamma_k+2)}{\gamma_\cdot(\gamma_\cdot+1)(\gamma_\cdot+2)(\gamma_\cdot+3)(\gamma_\cdot+4)(\gamma_\cdot+5)} \\
&= \frac{p_i[(1-\theta)p_i+\theta]p_j}{(1+\theta)} \\
&\quad \times \frac{p_k[(1-\theta)p_k+\theta][(1-\theta)p_k+2\theta)]}{(1+2\theta)(1+3\theta)(1+4\theta)}
\end{aligned}$$

PAIRS OF LOCI

All the theory to this point has been for a single genetic locus, whereas the human genome contains many thousands of loci. For purposes of human identification, it will obviously be better to use as much information, meaning as many loci, as possible. Some of the complications in going from one to many loci can be illustrated by considering two loci.

Table 4.9: Gamete probabilities from genotypes at two loci.

Genotype	Gametic arrangement	Gamete A_1B_1	A_1B_2	A_2B_1	A_2B_2
$A_1A_1B_1B_1$	A_1B_1/A_1B_1	1	0	0	0
$A_1A_1B_1B_2$	A_1B_1/A_1B_2	1/2	1/2	0	0
$A_1A_1B_2B_2$	A_1B_2/A_1B_2	0	1	0	0
$A_1A_2B_1B_1$	A_1B_1/A_2B_1	1/2	0	1/2	0
$A_1A_2B_1B_2$	A_1B_1/A_2B_2	$(1-c)/2$	$c/2$	$c/2$	$(1-c)/2$
$A_1A_2B_1B_2$	A_1B_2/A_2B_1	$c/2$	$(1-c)/2$	$(1-c)/2$	$c/2$
$A_1A_2B_2B_2$	A_1B_2/A_2B_2	0	1/2	0	1/2
$A_2A_2B_1B_1$	A_2B_1/A_2B_1	0	0	1	0
$A_2A_2B_1B_2$	A_2B_1/A_2B_2	0	0	1/2	1/2
$A_2A_2B_2B_2$	A_2B_2/A_2B_2	0	0	0	1

Consider locus **A** with alleles A_1 and A_2 and locus **B** with alleles B_1, and B_2. At locus **A** an individual may have genotype A_1A_1, A_1A_2, or A_2A_2, and at locus **B** the same individual may have genotype B_1B_1, B_1B_2, or B_2B_2. Taking both loci into account, there will therefore be nine different genotypes: $A_1A_1B_1B_1$, $A_1A_1B_1B_2$, ... , $A_1A_2B_2B_2$. These are displayed in Table 4.9. Individuals transmit one allele per locus to their children, so that there are four possible types of GAMETE or HAPLOTYPE: A_1B_1, A_1B_2, A_2B_1, and A_2B_2. Genotypes are formed by the union of maternal and paternal gametes.

If an individual is homozygous at one or both loci there is no ambiguity about the types of gametes that the individual received. For example, an $A_1A_1B_1B_2$ must have been formed from the union of A_1B_1 and A_1B_2 gametes, although it will not be known which one came from which parent. The gametic origins of the genotype can be emphasized by writing the genotype with a slash separating the gametes: A_1B_1/A_1B_2. All eight genotypes that are homozygous at one or both loci can be represented by one such gametic arrangement.

There is ambiguity with double heterozygotes: $A_1A_2B_1B_2$. These genotypes arise from two different gametic pairings: A_1B_1/A_2B_2 and A_1B_2/A_2B_1, as indicated in the middle element of the array in Table 4.9. It is not possible to distinguish these two only on the basis of the individual's genotype.

Linkage

For the transmission of gametes, there are three cases to consider. Doubly homozygous individuals received two copies of the same gamete and can transmit only that kind: $A_1A_1B_1B_1$ can transmit only A_1B_1 for example. Individuals who are homozygous at one locus and heterozygous at the other have received two different gametes, and can transmit each of these two with equal probability: $A_1A_1B_1B_2$ individuals transmit A_1B_1 and A_1B_2 equally often.

Because of the possibility of RECOMBINATION between loci, however, doubly heterozygous individuals can transmit four kinds of gametes–the two PARENTAL types that the individual received and the two RECOMBINANT types that are different from both parental types. We write the probability of recombination as c and note that each recombinant gamete is transmitted with the same probability of $c/2$ and each parental gamete is transmitted with probability $(1-c)/2$. Loci that are on different chromosomes, or that are far apart on the same chromosome, are said to be UNLINKED. For such loci, double heterozygotes produce all four types of gamete with equal probability. This corresponds to $c = 0.5$, and the mechanism by which recombination takes place usually ensures that $0 \leq c \leq 0.5$ for all pairs of loci. Loci that are (virtually) at the same position on a chromosome are said to be COMPLETELY LINKED and have zero recombination fraction, $c = 0$. For such loci, double heterozygotes can transmit only the two parental gametes–each with a probability of 0.5. To summarize this discussion, we show gametic probabilities in Table 4.9.

Linkage disequilibrium

Partly because of the linkage phenomenon, the probability with which an individual receives a specific **A** allele may be related to the probability with which it receives a specific allele at locus **B**. This phenomenon is known as LINKAGE DISEQUILIBRIUM, even though it can exist between the alleles at unlinked loci. Formally, the coefficient D_{AB} of linkage disequilibrium for alleles A and B at loci **A** and **B** is defined as the difference between the AB gamete proportion and the product of A and B allele proportions p_A, p_B:

$$D_{AB} \quad = \quad P_{AB} - p_A p_B$$

Even though the two alleles may have proportions that do not change over time, the proportion of the pair AB will change. In Box 4.8 we show how linkage disequilibrium decays as recombination rearranges pairs of alleles at

Table 4.10: Proportions of two-locus genotypes.

	Locus **B**			
Locus **A**	B_1B_1	B_1B_2	B_2B_2	Total
A_1A_1	$1 : P_{A_1B_1}^{A_1B_1}$	$2 : P_{A_1B_2}^{A_1B_1}$	$3 : P_{A_1B_2}^{A_1B_2}$	$P_{A_1A_1}$
A_1A_2	$4 : P_{A_2B_1}^{A_1B_1}$	$5 : P_{A_2B_1}^{A_1B_2} + P_{A_2B_2}^{A_1B_1}$	$6 : P_{A_2B_2}^{A_1B_2}$	$P_{A_1A_2}$
A_2A_2	$7 : P_{A_2B_1}^{A_2B_1}$	$8 : P_{A_2B_2}^{A_2B_1}$	$9 : P_{A_2B_2}^{A_2B_2}$	$P_{A_2A_2}$
Total	$P_{B_1B_1}$	$P_{B_1B_2}$	$P_{B_2B_2}$	1

different loci. The value D_{AB} in one generation changes to D'_{AB} in the next:

$$D'_{AB} = (1-c)D_{AB}$$

The linkage disequilibrium decays by a maximum amount of one-half each generation for unlinked genes, and this rate is quite fast. The derivation in Box 4.8 makes use of the notation displayed in Table 4.10 for two-locus genotype proportions.

Disequilibrium in Admixed Populations

One way in which linkage disequilibrium can be created is by the amalgamation of two populations, in the same way that the Wahlund effect generates Hardy-Weinberg disequilibrium. Even if two subpopulations are each in linkage equilibrium there may be linkage disequilibrium in the combined population. The amount of disequilibrium is proportional to the differences between subpopulations of allele proportions at the two loci. If the two subpopulations are indexed by α and β, and allele proportions are $p_{A\alpha}, p_{A\beta}$ for allele A and $p_{B\alpha}, p_{B\beta}$ for allele B, then in the whole population the disequilibrium is

$$D_{AB} = m_\alpha m_\beta [p_{A\alpha} - p_{A\beta}][p_{B\alpha} - p_{B\beta}]$$

where m_α and m_β are the proportions of the whole population in the two subpopulations. This result is derived in Box 4.9.

Box 4.8: Decay of linkage disequilibrium

The gamete proportion for any pair of alleles can be expressed as the product of allele proportions plus a linkage disequilibrium coefficient:

$$\begin{array}{ll} P_{A_1B_1} = p_{A_1}p_{B_1} + D_{A_1B_1} & P_{A_1B_2} = p_{A_1}p_{B_2} + D_{A_1B_2} \\ P_{A_2B_1} = p_{A_2}p_{B_1} + D_{A_2B_1} & P_{A_2B_2} = p_{A_2}p_{B_2} + D_{A_2B_2} \end{array}$$

When these gamete proportions are summed over all the alleles at one locus, they give the allele proportions at the other locus. If A_2 indicates the sum of all alleles other than A_1 at the **A** locus, and B_2 indicates the sum of all alleles other than B_1 at the **B** locus,

$$\begin{array}{ll} P_{A_1B_1} + P_{A_1B_2} = p_{A_1} & P_{A_2B_1} + P_{A_2B_2} = p_{A_2} \\ P_{A_1B_1} + P_{A_2B_1} = p_{B_1} & P_{A_1B_2} + P_{A_2B_2} = p_{B_2} \end{array}$$

and these equations imply that $D_{A_1B_1} = D_{A_2B_2}$ and $D_{A_1B_1} = -D_{A_1B_2} = -D_{A_2B_1}$. Summing the probabilities of gamete A_1B_1 in Table 4.9, and using the notation of Table 4.10, the proportion of A_1B_1 gametes in an offspring population is

$$\begin{aligned} P'_{A_1B_1} &= P_{A_1B_1}^{A_1B_1} + \frac{1}{2}(P_{A_1B_2}^{A_1B_1} + P_{A_2B_1}^{A_1B_1} + P_{A_2B_2}^{A_1B_1}) - \frac{c}{2}(P_{A_2B_2}^{A_1B_1} - P_{A_1B_2}^{A_2B_1}) \\ &= P_{A_1B_1} - c(P_{A_1B_1}P_{A_2B_2} - P_{A_1B_2}P_{A_2B_1}) \\ &= P_{A_1B_1} - cD_{A_1B_1} \end{aligned}$$

where the second step depends on random mating, implying that two-locus genotype proportions are the products of gamete proportions, and the third step depends upon the relation found above among the four linkage disequilibrium coefficients. Subtracting the (constant) product of allele proportions from each side of the transition equation gives

$$P'_{A_1B_1} - p_{A_1}p_{B_1} = P_{A_1B_1} - p_{A_1}p_{B_1} - cD_{A_1B_1}$$

so that

$$D'_{A_1B_1} = (1-c)D_{A_1B_1}$$

Box 4.9: Linkage disequilibrium in subdivided populations

The proportions $P_{AB_\alpha}, P_{AB_\beta}$ for AB gametes in subpopulations α, β reflect the linkage equilibrium within each one, and are therefore the products of the allele proportions in those subpopulations:

$$P_{AB\alpha} = p_{A\alpha}p_{B\alpha} \qquad P_{AB\beta} = p_{A\beta}p_{B\beta}$$

In the combined population, therefore:

$$\begin{aligned} P_{AB} &= m_\alpha P_{AB\alpha} + m_\beta P_{AB\beta} \\ &= m_\alpha p_{A\alpha} p_{B\alpha} + m_\beta p_{A\beta} p_{B\beta} \\ &= [m_\alpha p_{A\alpha} + m_\beta p_{A\beta}][m_\alpha p_{B\alpha} + m_\beta p_{B\beta}] \\ &\quad + m_\alpha m_\beta [p_{A\alpha} - p_{A\beta}][p_{B\alpha} - p_{B\beta}] \\ &= p_A p_B + m_\alpha m_\beta [p_{A\alpha} - p_{A\beta}][p_{B\alpha} - p_{B\beta}] \end{aligned}$$

which leads to the result for the linkage disequilibrium coefficient given in the text. Now suppose the population consists of a number of subpopulations, the ith being in proportion m_i. The linkage disequilibrium in the whole population, when each subpopulation is in linkage equilibrium, is

$$D_{AB} = \sum_i m_i (p_{Ai} - p_A)(p_{Bi} - p_B)$$

Here p_{Ai} and p_{Bi} are the allelic proportions in the ith subpopulation, and p_A and p_B are the overall proportions. Note that this linkage disequilibrium can be either positive or negative.

Exercise 4.10 Suppose a population consists of proportions 0.8 and 0.2 of two subpopulations, each of which is in Hardy-Weinberg and linkage equilibrium for genes **A** and **B** but with different proportions, 0.4 and 0.2 respectively, for allele A and different proportions 0.2 and 0.6 for allele B. What is the proportion of AB gametes and the coefficient D_{AB} of linkage disequilibrium in the combined population?

Multilocus Genotypic Proportions

At a single locus, we have seen that the Hardy-Weinberg law allows genotypic proportions to be expressed as products of allele proportions. The DNA profiles used in forensic science are almost always multilocus, and it is very convenient to be able to express profile proportions as products of the proportions of the alleles at each locus in the profile. Suppose the loci are **A, B, C,** ..., and that the profile of interest has alleles A_1 and A_2 at locus **A**, B_1 and B_2 at locus **B**, alleles C_1 and C_2 at locus **C** and so on. The PRODUCT RULE generalizes the Hardy-Weinberg law to

$$\Pr(A_1A_2B_1B_2C_1C_2\ldots) \quad = \quad 2^H p_{A_1}p_{A_2}p_{B_1}p_{B_2}p_{C_1}p_{C_2}\cdots$$

with H being the number of loci that are heterozygous in the profile. It is often stated that this result assumes both Hardy-Weinberg equilibrium and linkage disequilibrium, but in fact more is required. Linkage disequilibrium, as discussed here, is defined to refer to *pairs* of alleles–one per locus. The product rule requires complete independence among *all* alleles in the profile. In Chapter 5 we discuss how to address this issue from multilocus data.

SUMMARY

Forensic evidence involving DNA profiles cannot be interpreted fully unless the genetic nature of the evidence is taken into account. At the simplest level, this may amount to invoking assumptions of independence of the constituent alleles within a profile, but a much richer theory of population genetics is available to the forensic scientist. This theory enables such complications as population structure and relatedness to be folded into the the calculation of likelihood ratios that we believe are at the heart of interpreting scientific evidence.

Chapter 5

Statistical Genetics

INTRODUCTION

We now return to the single-stain transfer case we discussed in Chapter 2 and consider further evaluation of the likelihood ratio in Equation 2.5. Recall that we used G to denote the matching genotypes of the suspect and the crime stain.

$$\mathrm{L}R \;=\; \frac{1}{\Pr(G|H_d, I)}$$

The probability in the denominator was assigned the value P, the proportion of the population who have genotype G. Then

$$\mathrm{L}R \;=\; \frac{1}{P}$$

We are now going to combine the ideas developed in Chapters 3 and 4 to assign a value to P for a multilocus genotype using estimates of allele proportions.

One of the issues that has arisen in connection with DNA evidence statistics is that of the apparent disparity between the extreme smallness of a DNA genotype proportion and the limited size of the sample used for estimating the proportion. For example, how can a sample of 100 people provide credible probabilities that may be only one in a million?

To answer this question it is necessary to invoke both genetic and statistical reasoning. In this chapter we will explain such reasoning by means of an example based on the five-locus AmpliType^R ("Polymarker") profile shown in Table 5.1.

Table 5.1: An AmpliType^R profile.

Locus	Genotype
LDLR	*AB*
GYPA	*BB*
HBGG	*BC*
D7S8	*AB*
Gc	*BC*

ESTIMATING PROPORTIONS

We will see in Table 5.3 that there are three genotypes for *LDLR*, three for *GYPA*, six for *HBGG*, three for *D7S8* and six for *Gc*. This means that there are 972 five-locus genotypes that can potentially be revealed by the Polymarker system. The proportions of these genotypes will vary considerably, with some being relatively common and many being very rare. The simplest estimate of the proportion of any particular genotype in a population is just its proportion in a sample from that population, and this is the maximum likelihood estimate discussed in Chapter 3. Unless the sample is extremely large, however, most of these genotypes will not be found in the sample even if they are present in the population.

We can explain this last point as follows. Consider any one of the genotypes; as we have seen in Chapter 3, the number of individuals in a sample with that particular genotype has a binomial probability distribution. The two parameters of the distribution are the known sample size n and the unknown population proportion P of people with that type. We first use the binomial distribution to calculate the probability of there being *no* individuals in the sample with the genotype in question:

$$\begin{aligned} \Pr(\text{Z}\textit{ero copies of genotype in the sample}) &= \frac{n!}{0!n!}P^0(1-P)^n \\ &= (1-P)^n \end{aligned}$$

From this we can calculate the probability of there being one or more copies of this genotype in the sample:

$$\Pr(\text{A}\textit{t least one copy of the genotype in the sample}) = 1-(1-P)^n$$

In Table 5.2 we show, for different values of P, the approximate sample size required for this probability to be at least 0.95. Evidently, databases of

a few hundred, or even a few thousand, are not going to ensure that rare genotypes are seen.

Table 5.2: Approximate sample sizes n needed to have 0.95 probability of detecting a genotype for which the proportion is P.

P	n
1	1
0.1	30
0.01	300
0.001	3,000
0.000,1	30,000
0.000,01	300,000
0.000,001	3,000,000

The solution to the problem of estimating small multilocus genotype proportions rests on the assumption of independence between the constituent parts of the profile. If it is reasonable to assume independence between loci, a multilocus genotype proportion can be estimated reliably by multiplying together the constituent single-locus genotype proportions.

To illustrate the process, we use data collected by Cellmark Diagnostics from a sample of 103 people who were typed at the Polymarker loci. Table 5.3 is in five parts, each one corresponding to the locus shown in column 1. The second column shows the designations of the single-locus genotypes, and the numbers of people in the sample who have that genotype are shown in the third column. These counts are divided by the total of 103 to give the proportions shown in the fourth column. We know these sample proportions are estimates of the corresponding population proportions, and we can gain some idea of their precision by estimating their standard deviations. In Chapter 3 we saw that an estimated proportion $\hat{P}$, based on a sample of size n, has a probability distribution with a standard deviation of $\sqrt{P(1-P)/n}$, where P is the population proportion. We cannot use this equation directly because we don't know P, but statistical theory tells us that replacing P by $\hat{P}$ provides a good estimate of the standard deviation. These estimates, $\sqrt{\hat{P}(1-\hat{P})/n}$, are shown in the fifth column of Table 5.3.

The next thing we can do is recover the allele counts at each locus by the simple counting method described in Chapter 4. If these counts are divided by the total number of alleles, 2×103, we obtain sample allele proportions,

Table 5.3: Sample genotype proportions $\hat{P}$ (and standard deviations) for Polymarker loci.

		Observed values			Product estimates	
Locus	Genotype	Count	$\hat{P}$	(Std. dev.)	$\hat{P}$	(Std. dev.)
LDLR	*AA*	17	0.165	(0.037)	0.191	(0.030)
	AB	56	0.544	(0.049)	0.492	(0.009)
	BB	30	0.291	(0.045)	0.317	(0.039)
GYPA	*AA*	31	0.301	(0.045)	0.290	(0.037)
	AB	49	0.476	(0.049)	0.497	(0.005)
	BB	23	0.223	(0.041)	0.213	(0.032)
HBGG	*AA*	30	0.291	(0.045)	0.312	(0.039)
	AB	55	0.534	(0.049)	0.488	(0.010)
	AC	0	0.000	(0.000)	0.005	(0.005)
	BB	17	0.165	(0.037)	0.191	(0.030)
	BC	1	0.010	(0.010)	0.004	(0.004)
	CC	0	0.000	(0.000)	0.000	(0.000)
D7S8	*AA*	31	0.301	(0.045)	0.296	(0.038)
	AB	50	0.485	(0.049)	0.496	(0.006)
	BB	22	0.214	(0.040)	0.208	(0.032)
Gc	*AA*	4	0.039	(0.019)	0.064	(0.015)
	AB	11	0.107	(0.030)	0.100	(0.016)
	AC	33	0.320	(0.046)	0.277	(0.026)
	BB	8	0.078	(0.026)	0.040	(0.011)
	BC	14	0.136	(0.034)	0.218	(0.026)
	CC	33	0.320	(0.046)	0.301	(0.038)

Source: Cellmark Diagnostics.

Box 5.1: Standard deviation of estimated allele proportions

Allele proportions are linear combinations of genotype proportions. If the sample proportion of allele A_i is $\hat{p}_i$, and the sample proportion of genotype A_iA_j is $\hat{P}_{ij}$, then the variance of $\hat{p}_i$ is

$$\begin{aligned} \mathrm{V}ar(\hat{p}_i) &= \mathrm{V}ar(\hat{P}_{ii} + \frac{1}{2}\sum_{j\neq i}\hat{P}_{ij}) \\ &= \mathrm{V}ar(\hat{P}_{ii}) + \sum_{j\neq i}\mathrm{C}ov(\hat{P}_{ii},\hat{P}_{ij}) + \frac{1}{4}\sum_{j\neq i}\mathrm{V}ar(\hat{P}_{ij}) \\ &= \frac{1}{2n}(p_i + P_{ii} - 2p_i^2) \end{aligned}$$

where Cov denotes covariance, that is, the expected value of the product of two quantities minus the product of their expected values. When the Hardy-Weinberg relationship holds, so that $P_{ii} = p_i^2$, this reduces to

$$\mathrm{V}ar(\hat{p}_i) = \frac{1}{2n}p_i(1-p_i)$$

as expected for a sample from a binomial distribution $B(2n, p_i)$.

and these, in turn, are estimates of the population proportions. If we denote estimates of the proportions for alleles A and B in a system such as *LDLR* by $\hat{p}_A$, and $\hat{p}_B$, and if the three estimated genotype proportions are $\hat{P}_{AA}, \hat{P}_{AB}$, and $\hat{P}_{BB}$, then we have

$$\begin{aligned} \hat{p}_A &= \hat{P}_{AA} + \frac{1}{2}\hat{P}_{AB} \\ \hat{p}_B &= \hat{P}_{BB} + \frac{1}{2}\hat{P}_{AB} \end{aligned}$$

Allele proportion estimates are shown in Table 5.4, and once again the precision of the estimates can be gauged by the estimated standard deviations also displayed in this table. Theoretical expressions for these standard deviations are derived in Box 5.1.

Table 5.4: Sample allelic proportions $\hat{p}$ (and standard deviations) for Polymarker loci, using data in Table 5.3.

Locus	Allele	$\hat{p}$	(Std. dev.)
LDLR	*A*	0.437	(0.033)
	B	0.563	(0.033)
GYPA	*A*	0.539	(0.035)
	B	0.461	(0.035)
HBGG	*A*	0.558	(0.033)
	B	0.437	(0.033)
	C	0.005	(0.005)
D7S8	*A*	0.544	(0.035)
	B	0.456	(0.035)
Gc	*A*	0.252	(0.028)
	B	0.199	(0.031)
	C	0.549	(0.056)

Box 5.2: Standard deviation of estimated genotype proportions

The variance of products of sample proportions is not known exactly, but can be approximated by the following formulae (using Fisher's formula based on a Taylor series expansion: Weir 1996). If p_i is the proportion for allele A_i

$$\left.\begin{array}{rcl} \mathrm{V}ar(\hat{p}_i^2) & \approx & \frac{1}{2n} 4p_i^3(1-p_i) \\ \mathrm{V}ar(2\hat{p}_i\hat{p}_j) & \approx & \frac{1}{2n} 4p_ip_j(p_i+p_j-4p_ip_j) \end{array}\right\} \tag{5.2}$$

Sample allele proportions are substituted into these equations, and the square roots of the results are shown in Table 5.3.

THE PRODUCT RULE

So far we have made no assumptions about the conditions for independence of alleles within and between loci. If we can assume that the conditions for independence of alleles within loci exist in the sampled population, then we can use the Hardy-Weinberg formula to estimate population genotype proportions from the estimated allele proportions:

$$\left.\begin{array}{rcl} P_{AA} & \hat{=} & \hat{p}_A^2 \\ P_{AB} & \hat{=} & 2\hat{p}_A\hat{p}_B \\ P_{BB} & \hat{=} & \hat{p}_B^2 \end{array}\right\} \tag{5.1}$$

Here the symbol $\hat{=}$ means "is estimated by." These product estimates are displayed in column 6 of Table 5.3.

In the last column of Table 5.3 we give estimated standard deviations for the genotype proportions estimated as the products of allele proportions. The standard deviations are calculated by a method described in Box 5.2. Table 5.3 therefore shows the outcomes of two methods for estimating the population proportions of single-locus genotypes. The first estimate, in column 3, is from directly counting the number of times each genotype occurs in the sample. The second estimate, in column 5, is from applying the product rule to estimated allele proportions.

Which method of estimating genotype proportions is better? This is not a simple question and the answer depends on the meaning given to

"better" in this context. The counting method involves no independence assumptions, but the product rule method gives smaller standard deviations. Therefore the former estimates are intrinsically more accurate, although less precise, whereas the latter estimates are more precise but may be less accurate, depending on the validity of the independence assumption. These are very fine distinctions, and we would not expect them to be of any practical interest in forensic casework.

If independence of alleles is assumed both within and between loci, then the multilocus genotype proportion is estimated as the product of the sample proportions for all the alleles in the profile, multiplied by a factor of two for every locus that is heterozygous in the profile. This is the PRODUCT RULE. For the profile in Table 5.1 this product is 4.5×10^{-5} which is about two-thirds the value of the product of the five single-locus genotype proportions, and we expect it to have a smaller standard deviation than that of the product of genotype proportions.

Bayesian Approach

Bayesian methods can also be used to estimate genotype proportions as products of allele proportions on the basis of the Hardy-Weinberg law. We follow a method outlined by Balding (1995), and first look at the situation of estimating the proportion P_{12} of A_1A_2 genotypes. We have a particular interest in loci where the genotype may not appear in the sample, although the corresponding alleles do occur in the sample. If the Hardy-Weinberg law applies in the population, then $P_{12} = 2p_1p_2$, where p_1 and p_2 are the proportions for alleles A_1 and A_2.

For any heterozygote there are three allele categories: the two alleles of interest, and all other alleles, which we will collect together under the name A_3. Instead of a Beta prior distribution (Box 3.3), we use a Dirichlet prior for the three allele categories. In Box 5.3 we show that a Dirichlet prior combined with a multinomial likelihood results in a Dirichlet posterior distribution. Expected values of products of quantities that have a Dirichlet distribution may be found very easily, as shown in Chapter 4. We provide details in Box 5.3, and give a simple example here. In the absence of any other information, it is often the case that a uniform prior (a special case of the Dirichlet) is adopted for the allele proportions. If the counts of alleles A_1, A_2, and A_3 in a sample of $2n$ alleles are x_1, x_2, and x_3, then the posterior expectation of $2p_1p_2$ is

$$\mathcal{E}(2p_1p_2) \quad = \quad \frac{2(x_1+1)(x_2+1)}{(2n+3)(2n+4)}$$

and the corresponding result for homozygotes A_1A_1 is

$$\mathcal{E}(p_1^2) = \frac{(x_1+1)(x_1+2)}{(2n+3)(2n+4)}$$

Balding (1995) pointed out that, for large n, the heterozygote expression is very close to the maximum likelihood estimate that would result if one copy of the A_1A_2 genotype were added to the sample of n individuals. A similar interpretation for homozygotes would require that x_1 is also large. The posterior means provide nonzero estimates even if the allele counts in the sample are zero.

The posterior expected values of the squares of $2p_1p_2$ or p_1^2 also have relatively simple expressions:

$$\mathcal{E}(2p_1p_2)^2 = \frac{4(x_1+1)(x_1+2)(x_2+1)(x_2+2)}{(2n+3)(2n+4)(2n+5)(2n+6)}$$

$$\mathcal{E}(p_1^2)^2 = \frac{(x_1+1)(x_1+2)(x_1+3)(x_1+4)}{(2n+3)(2n+4)(2n+5)(2n+6)}$$

Under the proposition H_p, in the single-stain transfer case, there is only one person with the profile since the suspect is the person who left the crime sample, whereas under the proposition H_d there are two people, that is, the suspect and the offender. This led Balding (1995) to formulate the likelihood ratio as $\mathcal{E}(\phi)/\mathcal{E}(\phi^2)$ where ϕ is $2p_1p_2$ for heterozygote A_1A_2 and p_1^2 for homozygote A_1A_1. Using the posterior expectations from a uniform prior,

$$\frac{\mathcal{E}(2p_1p_2)}{\mathcal{E}(2p_1p_2)^2} = \frac{(2n+5)(2n+6)}{2(x_1+2)(x_2+2)}$$

$$\frac{\mathcal{E}(p_1^2)}{\mathcal{E}(p_1^2)^2} = \frac{(2n+5)(2n+6)}{(x_1+3)(x_1+4)}$$

For large n, the heterozygote expression is close to the reciprocal of the maximum likelihood estimate of the genotype proportion that would result from adding two copies of genotype A_1A_2 to the sample. A similar interpretation for the homozygote expression would require that x_1 is also large.

EFFECTS OF SUBPOPULATION DATA

When there is Hardy-Weinberg equilibrium within subpopulations, but variation in allele proportions among subpopulations, the allele proportions

Box 5.3: General Dirichlet prior
For more than two categories, we can consider the Dirichlet distribution $D(\gamma_1, \gamma_2, \ldots)$ for proportions $p_1, p_2, \ldots$ instead of the Beta distribution $Be(\alpha, \beta)$ for $p, 1-p$. The pdf is

$$f(p_1, p_2, \ldots) = \frac{\Gamma(\gamma_1 + \gamma_2 + \ldots)}{\Gamma(\gamma_1)\Gamma(\gamma_2)\ldots} p_1^{\gamma_1 - 1} p_2^{\gamma_2 - 1} \ldots, \quad 0 \leq p_i \leq 1$$

When $\gamma_i = 1$ for all i, the Dirichlet reduces to a uniform distribution.
For a dataset of $2n$ alleles, with counts $(x_1, x_2, \ldots)$, Bayes' theorem leads to a Dirichlet posterior distribution:

$$\begin{aligned} \pi(p_1, p_2, \ldots | x_1, x_2, \ldots) &= \frac{\Gamma(\gamma_\cdot + 2n)}{\Gamma(\gamma_1 + x_1)\Gamma(\gamma_2 + x_2)\ldots} \\ &\quad \times p_1^{\gamma_1 + x_1 - 1} p_2^{\gamma_2 + x_2 - 1} \ldots \end{aligned}$$

where $\gamma_\cdot = \gamma_1 + \gamma_2 + \ldots, 2n = x_1 + x_2 + \ldots$ Therefore, the Dirichlet is a CONJUGATE DISTRIBUTION for the multinomial.
For allele A_i, after having seen x_i copies in a sample of size $2n$, the allele probability becomes

$$\begin{aligned} \mathcal{E}(p_i | x_i) &= \frac{\Gamma(\gamma_i + x_i + 1)}{\Gamma(\gamma_i + x_i)} \frac{\Gamma(\gamma_\cdot + 2n)}{\Gamma(\gamma_\cdot + 2n + 1)} \\ &= \frac{\gamma_i + x_i}{\gamma_\cdot + 2n} \end{aligned}$$

which is between the prior probability $\gamma_i/\gamma_\cdot$ and the sample value x_i/n. More generally, the posterior mean of a product $p_1^{t_1} p_2^{t_2} \ldots$ is

$$\begin{aligned} \mathcal{E}(p_1^{t_1} p_2^{t_2} \ldots | x_1, x_2, \ldots) &= \frac{\Gamma(\gamma_\cdot + 2n + t)}{\Gamma(\gamma_\cdot + 2n)} \frac{\prod_i \Gamma(\gamma_i + x_i + t_i)}{\prod_i \Gamma(\gamma_i + x_i)} \\ &= \frac{\prod_i \prod_{j=0}^{t_i - 1} (\gamma_i + x_i + j)}{\prod_{j=0}^{t-1} (\gamma_\cdot + 2n + j)} \end{aligned}$$

providing each t_i is greater than or equal to 1.

within a subpopulation may be described by a Dirichlet distribution with parameters $\gamma_i = (1-\theta)p_i/\theta$. As we discussed in Chapter 4, the quantity θ describes the relationship of alleles within the subpopulation relative to that between subpopulations, and p_i is the reference proportion for allele A_i (i.e., the proportion averaged over all subpopulations). The expected values of the genotype proportions are

$$\begin{aligned} \Pr(A_iA_i) &= p_i[(1-\theta)p_i + \theta] \\ \Pr(A_iA_j) &= 2(1-\theta)p_ip_j, \;\; i \neq j \end{aligned}$$

and these are taken to apply to any subpopulation in the absence of data from the subpopulation. What would be the effect of having a sample of alleles from the subpopulation?

We can invoke the reasoning of Box 5.3, with the Dirichlet parameter values $\gamma_i = (1-\theta)p_i/\theta$. In Box 5.4 we show that posterior expected values of the genotype proportions, for large samples, are

$$\begin{aligned} \Pr(A_iA_i) &= \tilde{p}_i^2 \\ \Pr(A_iA_j) &= 2\tilde{p}_i\tilde{p}_j, \;\; i \neq j \end{aligned}$$

where the $\tilde{p}_i$ and $\tilde{p}_j$ terms are allele proportions in the sample from the particular relevant subpopulation. We have used a tilde instead of a hat to distinguish the subpopulation sample proportion from the proportion in a sample from the whole population. This result is just as we would hope, and it serves to emphasize that the Dirichlet distribution formulation, on which Equation 4.23 is based, assumes Hardy-Weinberg within subpopulations.

In Chapter 4 we also gave derivations for genotype proportions conditional on the genotype having been seen once already. For individuals X and Y Equations 4.20 are

$$\begin{aligned} \Pr(G_X = A_iA_i|G_Y = A_iA_i) &= \frac{[2\theta + (1-\theta)p_i][3\theta + (1-\theta)p_i]}{(1+\theta)(1+2\theta)} \\ \Pr(G_X = A_iA_j|G_Y = A_iA_j) &= \frac{2[\theta + (1-\theta)p_i][\theta + (1-\theta)p_j]}{(1+\theta)(1+2\theta)} \end{aligned}$$

Once again, p_i is the allele reference proportion, and θ describes variation among subpopulations. Using the same kind of argument as in Box 5.4, we can show that if data are available from the relevant subpopulation, then these genotype probabilities also reduce to

$$\begin{aligned} \Pr(G_X = A_iA_i|G_Y = A_iA_i) &= \tilde{p}_i^2 \\ \Pr(G_X = A_iA_j|G_Y = A_iA_j) &= 2\tilde{p}_i\tilde{p}_j, \;\; i \neq j \end{aligned}$$

when $\tilde{p}_i$ is the sample proportion in the subpopulation. The availability of data from the subpopulation removes the need for the θ formulation.

CONFIDENCE INTERVALS

Whatever estimate we derive for the population proportion of a particular genotype, we recognize that it is subject to uncertainty. There are several sources of uncertainty, as we discuss later, but one refers to the fact that we must base our estimate on a sample of limited size. In Chapter 3 we met one approach to dealing with uncertainty created by sampling effects: the use of confidence intervals.

In Chapter 3 we showed that the binomial distribution $B(n, P)$ for the sample proportion $\hat{P}$ is well approximated by the normal distribution $N(P, P(1-P)/n)$. This requires a large sample size n and a population proportion P that is not too far from 0.5. If we substitute the sample proportion $\hat{P}$ for P, then the normal distribution provides a 95% confidence interval for P as

$$\hat{P} \quad \pm \quad 1.96\sqrt{\hat{P}(1-\hat{P})/n} \tag{5.3}$$

The justification for use of such an interval is that, provided the conditions for the various approximations are valid, in the long run the intervals will contain the true value of P with probability 0.95. In the column headed "Normal" of Table 5.5 we show the confidence intervals calculated from Equation 5.3 for the *HBGG* genotype proportions.

We must emphasize the limited usefulness of the single-locus genotype proportion confidence intervals. They were constructed on the basis of the normal approximation to the binomial, which is unlikely to be good for small genotype proportions. The symmetry of the normal confidence intervals means that the lower limit is often negative and has to be truncated at zero. Of more importance, however, is the fact that we are interested in multilocus profiles and not single-locus genotypes. Normal-theory intervals can be constructed for multiple loci, but it is not clear that they are valid in the present situation. An alternative way of calculating confidence intervals is provided by BOOTSTRAPPING (Efron 1982; Efron and Tibshirani 1993). This method is conceptually simple, and makes fewer assumptions, although it does impose a computational burden.

Bootstrapping is a method for simulating the process of taking a sample from a population. When we discussed the binomial and multinomial distributions in Chapter 3, we used the example of drawing balls from an urn, and we were careful to stipulate that after each ball was sampled it was replaced

Box 5.4: Effects of subpopulation data
Suppose a sample of size n is available from the relevant subpopulation, and that the sample contains x_i copies of allele A_i, where $\sum_i x_i = 2n$, so that the sample allele proportions are $\tilde{p}_i = x_i/(2n)$. If K is the number of ways of arranging the $2n$ alleles among the n genotypes, i.e., the ordering of distinct genotypes plus ordering the alleles within heterozygotes, then the probability of the sample plus the crime stain genotype is

$$\begin{aligned}
\Pr(A_iA_i, sample) &= K\frac{\Gamma(\gamma_.)}{\Gamma(\gamma_. + 2n + 2)}\frac{\Gamma(\gamma_i + x_i + 2)}{\Gamma(\gamma_i)}\prod_{k\neq i}\frac{\Gamma(\gamma_k + x_k)}{\Gamma(\gamma_k)} \\
\Pr(A_iA_j, sample) &= 2K\frac{\Gamma(\gamma_.)}{\Gamma(\gamma_. + 2n + 2)}\frac{\Gamma(\gamma_i + x_i + 1)}{\Gamma(\gamma_i)}\frac{\Gamma(\gamma_j + x_j + 1)}{\Gamma(\gamma_j)} \\
&\quad \times \prod_{k\neq i,j}\frac{\Gamma(\gamma_k + x_k)}{\Gamma(\gamma_k)}, \quad i \neq j
\end{aligned}$$

whereas the probability of the sample is

$$\Pr(sample) = K\frac{\Gamma(\gamma_.)}{\Gamma(\gamma_. + 2n)}\prod_{k}\frac{\Gamma(\gamma_k + x_k)}{\Gamma(\gamma_k)}$$

Therefore, the crime stain genotype probabilities, conditional on the information in the sample, are

$$\begin{aligned}
\Pr(A_iA_i|sample) &= \frac{\Gamma(\gamma_.)}{\Gamma(\gamma_. + 2n + 2)}\frac{\Gamma(\gamma_i + x_i + 2)}{\Gamma(\gamma_i)} \\
&= \frac{\left[\theta\tilde{p}_i + \frac{(1-\theta)p_i + \theta}{2n}\right]\left[\theta\tilde{p}_i + \frac{(1-\theta)p_i}{2n}\right]}{\left[\theta + \frac{1-\theta}{2n}\right]\left[\theta + \frac{1}{2n}\right]}
\end{aligned}$$

and

$$\begin{aligned}
\Pr(A_iA_j|sample) &= 2\frac{\Gamma(\gamma_.)}{\Gamma(\gamma_. + 2n + 2)}\frac{\Gamma(\gamma_i + x_i + 1)}{\Gamma(\gamma_i)}\frac{\Gamma(\gamma_j + x_j + 1)}{\Gamma(\gamma_j)} \\
&= 2\frac{\left[\theta\tilde{p}_i + \frac{(1-\theta)p_i}{2n}\right]}{\left[\theta + \frac{1}{2n}\right]}\frac{\left[\theta\tilde{p}_j + \frac{(1-\theta)p_j}{2n}\right]}{\left[\theta + \frac{1-\theta}{2n}\right]}
\end{aligned}$$

As the sample size increases, these probabilities reduce to products of the sample allele proportions.

Table 5.5: *HBGG* genotype 95% confidence intervals, using data in Table 5.3.

Genotype		Estimate	Confidence intervals	
			Normal	Bootstrap
AA	obs.	0.291	(0.204, 0.379)	(0.214, 0.388)
	exp.	0.312	(0.236, 0.387)	(0.245, 0.386)
AB	obs.	0.534	(0.438, 0.630)	(0.437, 0.631)
	exp.	0.488	(0.469, 0.507)	(0.463, 0.500)
AC	obs.	0.000	(0.000, 0.000)	(0.000, 0.000)
	exp.	0.005	(-0.005, 0.016)	(0.000, 0.017)
BB	obs.	0.165	(0.093, 0.237)	(0.097, 0.243)
	exp.	0.191	(0.132, 0.250)	(0.140, 0.255)
BC	obs.	0.010	(-0.009, 0.029)	(0.000, 0.029)
	exp.	0.004	(-0.004, 0.013)	(0.000, 0.014)
CC	obs.	0.000	(0.000, 0.000)	(0.000, 0.000)
	exp.	0.000	(0.000, 0.000)	(0.000, 0.000)

to ensure that the proportions of each type of ball remained constant. A bootstrap sample is created by sampling from the original sample *with replacement*, so each individual may be chosen 0, 1, 2, 3, or even more times. It is usual practice to make the bootstrap sample the same size as the original sample. Once we have a bootstrap sample we can estimate anew the parameter of interest–a multilocus genotype proportion in this case. Furthermore, if we create a large number of bootstrap samples, say 1,000 or more, then we gain an impression of the variability of our estimator. The bootstrap confidence interval follows naturally from this process, by recording the range of values within which the middle 95% of the bootstrap estimates fall (i.e., numbers 25 through 975 for an ordered set of 1,000). Clearly we could not perform this process by hand, but it is a simple computer task. The method is illustrated for single-locus genotypes in Box 5.5.

In the last column of Table 5.5 we show 95% confidence intervals for the *HBGG* genotypes, created by resampling the original sample of 103 individuals 1,000 times. These confidence intervals are not greatly different from the normal-theory intervals, suggesting that the normal approximation works quite well even when P values are considerably smaller than 0.5.

The method for calculating a normal confidence interval for the full five-locus profile proportion is shown in Box 5.6. Applying that procedure to our sample of 103 individuals leads to the interval (0, 0.000145). We found a bootstrap confidence interval, based on 1,000 resamples, to be the wider interval (0, 0.000187). The difference between the upper limits from the two methods becomes greater as the proportion of interest becomes smaller. We believe that the bootstrap intervals are more appropriate.

We need to note the meaning of confidence intervals, whether they are calculated from normal theory or by bootstrapping, and sound a note of caution if they are to be used in court. In the first place, a confidence interval is not a probability statement about the unknown proportion. Instead it is a statement about all the intervals that are calculated from all the possible samples from the population–95% of these intervals are expected to enclose the true proportion. This distinction may appear subtle or confusing and may be missed by a jury. However, the forensic scientist giving evidence in terms of confidence intervals should have no doubts as to their meaning. Secondly, the range of values in the interval reflects the uncertainty in the estimated profile proportion due to the use of a sample rather than the whole population. This range does not refer to any other source of uncertainty such as the relevance of the sample or population structure. Population structure effects, for example, are accommodated by the approach illustrated in Equations 4.20. Confidence intervals cannot refer to any errors in sample

Box 5.5: Bootstrap procedure

Suppose a sample of size 10, numbered $0, 1, \ldots, 9$, has genotypes

0	1	2	3	4	5	6	7	8	9
AB	AB	BB	AA	AA	AB	AA	BB	AA	AB

The sample proportions of the two alleles are $\hat{p}_A = 0.6$, $\hat{p}_B = 0.4$, and the estimated proportion of the AB genotype is $2\hat{p}_A\hat{p}_B = 0.48$.
A bootstrap sample is constructed by drawing, with replacement, ten individuals from the sample. This could be done using random numbers like those in Table A.3. Such a set is 30246 86149, and the new sample is

0	1	2	3	4	5	6	7	8	9
AA	AB	BB	AA	AA	AA	AA	AB	AA	AB

Individual 6 in the original sample is represented twice in the new sample and individual 5 is not represented at all. The allele proportions in the new sample are $\hat{p}_A = 0.75$, $\hat{p}_B = 0.25$, and the new heterozygote proportion estimate is $2\hat{p}_A\hat{p}_B = 0.375$.
The process is repeated many times, always drawing from the original sample. The set of resulting estimates (of the heterozygote proportion) provides an estimate of the distribution of these proportion estimates.

Box 5.6: Multilocus proportion confidence intervals

To generate a normal-theory confidence interval, it is necessary to extend Equation 5.3 to multiple loci. If P_l is the genotypic proportion for the lth locus in the profile, then the variance of the product $\prod_l \hat{P}_l$ of the sample proportions $\hat{P}_l$ is approximately

$$\mathrm{V}ar(\prod_l \hat{P}_l) \quad = \quad (\prod_l P_l)^2 \left(\prod_l \left[1 + \frac{\mathrm{V}ar(\hat{P}_l)}{P_l^2} \right] - 1 \right)$$

(Goodman 1960; Chakraborty et al. 1993). This variance was used to construct a normal-theory confidence interval for the Polymarker profile in Table 5.1. The interval is $(-0.000049, 0.000145)$, which is truncated to $(0, 0.000145)$.

preparation or genotyping. The width of the interval indicates the uncertainty brought about by the size of the sample. Larger samples will lead to smaller intervals. In spite of their limited value, the fact that confidence intervals for a profile proportion P may be as wide as $(\hat{P}/10, 10\hat{P})$ signals a clear need for caution. Very small proportions cannot be estimated with high precision from small samples.

INDEPENDENCE TESTING

We have seen that the product rule is correct only when all the alleles in a profile are independent in the population of interest. We have also seen that the conditions leading to independence are not satisfied in real-world populations, so the assumptions of independence can only be convenient approximations. In practice, we expect that the departures from independence are too small to have practical consequences. One way of studying whether this is the case is to use classical statistical hypothesis testing, even though a failure to detect departures from independence by these tests does not mean there are no departures. Instead, we hope that the tests would detect departures large enough to have practical consequences.

From our knowledge of the biology of the genetic markers used for identification, we expect the most likely type of departure from independence is

Table 5.6: Goodness-of-fit test for Hardy-Weinberg proportions for *LDLR* data in Table 5.3.

Genotype	Observed number o	Expected number e	$(o-e)$	$(o-e)^2/e$
AA	17	19.66	−2.66	0.36
AB	56	50.68	5.32	0.56
BB	30	32.66	−2.66	0.22
Total	103	103	0.00	1.14

for the two alleles at one locus, and this is most likely to be a consequence of population structure. We mentioned this in Chapter 4.

One-Locus Testing

Examining the independence of alleles at one locus is known as HARDY-WEINBERG TESTING, and many tests have been proposed (reviewed in Weir 1996). We will examine the traditional goodness-of-fit test here, as well as the more powerful exact test.

Goodness-of-fit test. In Table 5.3 we showed some observed genotypic proportions for the Polymarker system along with the proportions expected under the Hardy-Weinberg hypothesis (the one-locus product rule). At that point we commented that the two sets of proportions appeared to be in good agreement, even though they weren't exactly the same for any of the 21 genotypes. The goodness-of-fit test gives us one method for quantifying the extent to which observed and expected proportions agree or disagree. The test involves calculating a test statistic and comparing this to the probability distribution known to hold when the null hypothesis is true. In this case, the null hypothesis is that the Hardy-Weinberg relation holds.

We first described the chi-square goodness-of-fit test in Chapter 3 and illustrated it by means of examples for a roulette wheel. Now we apply it to the *LDLR* data in Table 5.3, and we lay out the calculations in Table 5.6. The first step is to determine the allele counts by using the relationship we described in Chapter 4: $n_A = 90, n_B = 116$. The estimated allele proportions are therefore $(90/206) = 0.437, (116/206) = 0.563$ as shown in

Table 5.4.

The test is based on the formulation of a null hypothesis, here that the population from which the sample was drawn has genotypes in Hardy-Weinberg proportions. This is equivalent to independence between alleles within a locus. We calculate the numbers of the three genotypes that we expect to see in a sample of size 103 if this null hypothesis is true:

$$\begin{aligned} e_{AA} &= 103 \times (90/206)^2 = 19.66 \\ e_{AB} &= 206 \times (90/206) \times (116/206) = 50.68 \\ e_{BB} &= 103 \times (116/206)^2 = 32.66 \end{aligned}$$

These expected numbers are shown in the third column of Table 5.6, and they are equivalent to the "Product estimates" in the fifth column of Table 5.3. Evidently, the word "expected" here has a different meaning than when we were talking about expected values in Chapter 3. Strictly, the e terms should be called "estimates assuming the hypothesis is true," but the single word "expected" is conventional in this context.

Now we can calculate the (Observed − Expected)^2/Expected terms that we met in Chapter 3. We show them in the fourth column of Table 5.6 and the sum of these terms is $X^2 = 1.14$:

$$X^2 = \sum_{genotypes} \frac{(\mathrm{O}bserved - Expected)^2}{\mathrm{E}xpected} = 1.14$$

To relate this number to the chi-square distribution, we need to know the degrees of freedom. As the name suggests, this is the number of categories (genotypes) whose expected numbers we are free to choose without reference to expected numbers in the other categories. In this example, suppose we start by choosing the expected number for the AA genotype, using the allele proportion $\hat{p}_A$ found in the sample. With this expected number fixed, the number for AB is constrained by the need for the AA number and half the AB number to sum to the number of A alleles. In other words, the number for AA determines the number for AB, and together they determine the number for BB because all three numbers must sum to the sample size. The degrees of freedom for the Hardy-Weinberg test at a locus with two alleles is one. More generally, the degrees of freedom for a locus with m alleles is $m(m-1)/2$, or the number of genotypes minus the number of alleles.

Recall from Chapter 3 that the 5% upper tail for the chi-square distribution with 1 df is delimited by the value of 3.84. Test statistics greater than 3.84 lie among the least probable 5% of values expected if the null

hypothesis is true. For this *LDLR* sample, the test statistic is much less than 3.84, and we do not reject the Hardy-Weinberg hypothesis.

It is important to note that the test must be performed on counts, not on proportions or percentages. Suppose the data were 90 AA, 0 AB, and 10 BB. The absence of AB genotypes is a strong indication of departures from Hardy-Weinberg equilibrium, and the test statistic has the value of 100:

$$X^2 = \frac{(90-81.0)^2}{81.0} + \frac{(0-18.0)^2}{18.0} + \frac{(10-1.0)^2}{1.0} = 100$$

If the statistic had been calculated incorrectly with proportions (0.90, 0.00, 0.10 observed and 0.81, 0.18, 0.01 expected) rather than counts, however, it would have the value of 1 and would not lead to rejection of the Hardy-Weinberg hypothesis.

We explained in Chapter 3 that the chi-square goodness-of-fit test becomes unreliable when one or more of the expected counts are small. It is conventional to require that all expected counts are at least five, although this ad-hoc rule can be relaxed. For loci with many alleles, however, it is common to find small expected counts even for large sample sizes, and for this reason we prefer to use exact tests.

Exercise 5.1 Perform goodness-of-fit tests for Hardy-Weinberg for: (a) *GYPA*; (b) *HBGG*; (c) *D7S8*; and (d) *Gc* using the data in Table 5.3.

Exact Tests. In Chapter 3 we explained that the exact test is based on the idea of using the multinomial distribution to calculate the probability of the observed data given that the null hypothesis is true. The hypothesis is rejected if this probability belongs to the set of smallest (5%) of possible values. This is in contrast to the goodness-of-fit test, which rejects the hypothesis when the test statistic is larger than expected if the hypothesis is true.

For a locus such as *LDLR* with two alleles A and B, the probability needed is that of the genotype counts in the sample conditional on the allele counts and conditional on the Hardy-Weinberg hypothesis being true. In Box 5.7 we derive this probability:

$$\Pr(n_{AA}, n_{AB}, n_{BB} | n_A, n_B) = \frac{n! n_A! n_B! 2^{n_{AB}}}{(2n)! n_{AA}! n_{AB}! n_{BB}!} \qquad (5.4)$$

This equation looks a little forbidding at first sight, and it is usually evaluated by computer, but we can illustrate its use for the *LDLR* data in

Table 5.3:

$$\Pr(n_{AA}, n_{AB}, n_{BB}|n_A, n_B) = \frac{103!90!116!2^{56}}{206!17!56!30!} = 0.0958$$

For loci with more than two alleles, there is a natural extension of Equation 5.4. We index the genotypes by g and the alleles by a, and write H for the number of individuals in the sample who are heterozygous at this locus. The equation becomes

$$\Pr(\{n_g\}|\{n_a\}) = \frac{n!2^H \prod_a n_a!}{(2n)! \prod_g n_g!}$$

where $\{n_g\}$ refers to the collection of genotype counts and $\{n_a\}$ refers to the collection of allele counts.

At this point we must be careful to avoid a possible source of confusion. When we performed the goodness-of-fit test we referred to the chi-square distribution to find the probability of the calculated test statistic *or a greater value* if the null hypothesis was true. This tail area probability is also called the significance level. It corresponds to unlikely values if the hypothesis is true, and these unlikely values are the *largest* values. When we perform the exact test the test statistic itself is a probability, but the statistic does not give the tail probability. The tail probability is the sum of the probabilities for all the outcomes that are as probable as or less probable than the observed outcome, i.e., we add together the smallest probabilities.

How do we determine if the exact-test probability indicates that the sample belongs to the set of unusual outcomes if the Hardy-Weinberg hypothesis is true? We need the total probability of all the possible outcomes that are as probable as or are less probable than the observed outcome. This total is the significance level. We can either quote this value for any set of genotypes, or we can say the hypothesis is rejected at the α% level if the significance level is less than α. Because the observed outcome belongs to the set of outcomes that determine the significance level, its probability is less than (or equal to) the significance level. In our *LDLR* example the probability of the data is already bigger than the conventional 0.05 level for hypothesis rejection, so the significance level must also be bigger than 0.05.

All 46 possible sets of genotypes for $n_A = 90, n_B = 116$ are shown in Table 5.7, along with their probabilities calculated from Equation 5.4, which assumes the Hardy-Weinberg hypothesis is true. Notice only five of the 46 have probabilities greater than the 0.0958 value for our data. The sum of the probabilities for the remaining 41 sets (which includes the data) is 0.3239. So, for these data, the significance level or P-value is 0.3239.

Box 5.7: Probabilities needed for the exact test of the Hardy-Weinberg hypothesis

Recall that the multinomial probability for genotype counts n_{AA}, n_{AB}, and n_{BB} for a locus with alleles A and B is

$$\Pr(n_{AA}, n_{AB}, n_{BB}) = \frac{n!}{n_{AA}!n_{AB}!n_{BB}!}(p_{AA})^{n_{AA}}(p_{AB})^{n_{AB}}(p_{BB})^{n_{BB}}$$

Under the Hardy-Weinberg hypothesis, genotype proportions are replaced by products of allele proportions, and the probability can be written as

$$\Pr(n_{AA}, n_{AB}, n_{BB}, n_A, n_B) = \frac{n!}{n_{AA}!n_{AB}!n_{BB}!}2^{n_{AB}}(p_A)^{n_A}(p_B)^{n_B} \quad (5.5)$$

where n_A and n_B are the two allele counts. The difficulty with this expression is that we don't know the population allele proportions. One way around this is to work with the probability of the genotype counts conditional on the allele counts. We ask whether the arrangement of n_A alleles of type A and n_B alleles of type B into genotype counts n_{AA}, n_{AB}, and n_{BB} falls among the most unlikely arrangements if the Hardy-Weinberg relation holds. This has been found to lead to satisfactory Hardy-Weinberg tests (Maiste and Weir 1995).
From the third law of probability,

$$\Pr(n_{AA}, n_{AB}, n_{BB}|n_A, n_B) = \frac{\Pr(n_{AA}, n_{AB}, n_{BB}, n_A, n_B)}{\Pr(n_A, n_B)} \quad (5.6)$$

What is the probability of the allele counts? Under the Hardy-Weinberg hypothesis, alleles are independent, so a sample of n genotypes is equivalent to a sample of $2n$ alleles. The binomial distribution holds for the two alleles:

$$\Pr(n_A, n_B) = \frac{(2n)!}{n_A!n_B!}(p_A)^{n_A}(p_B)^{n_B} \quad (5.7)$$

Putting Equations 5.5, 5.6, and 5.7 together provides the conditional probability of the genotype counts if Hardy-Weinberg holds:

$$\Pr(n_{AA}, n_{AB}, n_{BB}|n_A, n_B) = \frac{n!n_A!n_B!2^{n_{AB}}}{(2n)!n_{AA}!n_{AB}!n_{BB}!}$$

When this probability is used for an exact test with permutation-based significance levels, the value obtained for a data set is compared to values obtained for data sets formed by permuting alleles among genotypes. In each of the permuted sets, the samples sizes $(n, 2n)$ remain the same, as do the allele counts (n_a). Comparisons can therefore be restricted to the ratio $2^{n_{AB}}/(n_{AA}!n_{AB}!n_{BB}!)$.

Table 5.7: All possible samples with $n_A = 90, n_B = 116$, together with probabilities calculated assuming the Hardy-Weinberg hypothesis is true.

n_{AA}	n_{AB}	n_{BB}	Prob.	Cum. prob.	n_{AA}	n_{AB}	n_{BB}	Prob.	Cum. prob.
45	0	58	0.0000	0.0000	31	28	44	0.0000	0.0000
44	2	57	0.0000	0.0000	9	72	22	0.0000	0.0000
43	4	56	0.0000	0.0000	30	30	43	0.0000	0.0000
42	6	55	0.0000	0.0000	10	70	23	0.0001	0.0001
41	8	54	0.0000	0.0000	29	32	42	0.0001	0.0003
40	10	53	0.0000	0.0000	11	68	24	0.0004	0.0007
0	90	13	0.0000	0.0000	28	34	41	0.0005	0.0012
39	12	52	0.0000	0.0000	12	66	25	0.0016	0.0028
1	88	14	0.0000	0.0000	27	36	40	0.0019	0.0047
38	14	51	0.0000	0.0000	13	64	26	0.0052	0.0098
2	86	15	0.0000	0.0000	26	38	39	0.0057	0.0155
37	16	50	0.0000	0.0000	14	62	27	0.0138	0.0293
3	84	16	0.0000	0.0000	25	40	38	0.0148	0.0441
36	18	49	0.0000	0.0000	15	60	28	0.0310	0.0751
4	82	17	0.0000	0.0000	24	42	37	0.0327	0.1078
35	20	48	0.0000	0.0000	16	58	29	0.0591	0.1668
5	80	18	0.0000	0.0000	23	44	36	0.0613	0.2282
34	22	47	0.0000	0.0000	17	56	30	0.0958	0.3239
6	78	19	0.0000	0.0000	22	46	35	0.0982	0.4221
33	24	46	0.0000	0.0000	18	54	31	0.1321	0.5542
7	76	20	0.0000	0.0000	21	48	34	0.1340	0.6882
32	26	45	0.0000	0.0000	19	52	32	0.1555	0.8438
8	74	21	0.0000	0.0000	20	50	33	0.1562	1.0000

Permutation-Based Significance Levels

For a locus with only two alleles, it is not difficult to examine all possible genotypic arrays for a given allelic array as shown in Table 5.7, particularly when the sample sizes are not too large. For loci with many alleles, however, the number of genotypic arrays is too large to handle even on a computer. In those situations we employ PERMUTATION procedures. Instead of examining all possible genotypic arrays, we choose a random sample of the arrays by permuting (or shuffling) the alleles.

The process can be visualized as one of constructing a deck of cards, one for each of the n individuals in the sample. One side of each card is labeled with the two alleles observed for that individual, and then the card is torn in half between the two labels. This results in a deck of $2n$ cards, each card now showing just one allele label. Tearing the cards corresponds to breaking whatever association there is between alleles within individuals. The deck is then shuffled and dealt into n pairs. These pairs provide the genotypes for a new array in which the allelic array is the same as for the original data, and this new genotypic array is a random choice from all possible arrays. The exact test probability is evaluated for the new array and is compared to that for the original data. If m permutations are performed, the number of permuted arrays that are as probable, or less probable, than the original data has the binomial distribution $B(m, P)$. Therefore the proportion of permuted arrays with a probability no more than that for the data gives an estimate of the P-value.

To be 95% sure of achieving estimates within 0.01 of the P-value that would result from examining all genotypic arrays, it is necessary to perform $m = 10{,}000$ permutations. This follows from properties of the binomial distribution. Invoking the normal approximation to the binomial distribution allows us to say with 95% probability that estimated values of P lie within 0.01 of the true value when

$$0.01 \quad \geq \quad 1.96\sqrt{P(1-P)/m}$$

Because the right-hand size is largest when $p = 0.50$, we find that $m \geq 10{,}000$. If interest was centered on values around $p = 0.05$, however, we have

$$\begin{aligned} 0.01 &\geq 1.96\sqrt{0.05 \times 0.95/m} \\ m &\geq 2,000 \end{aligned}$$

Even the 10,000 value is still substantially less computing than examining all arrays. In practice, computing speeds are now sufficiently high that

performing 10,000 permutations does not take undue time. However, it is possible to reduce the time by keeping a running total of permuted data sets that are less probable, under the independence hypothesis, than the original data. Suppose 1,000 data sets were generated and 600 were found to be less probable. The P-value at that stage is 0.6, and it could not drop below 0.05 even if the next 9,000 sets gave a more probable data set–there would be no point, therefore, in performing 10,000 permutations.

When we perform an exact test by enumerating all outcomes, we report the sum of the probabilities of all outcomes with the same or smaller probabilities than that of the data. When we cannot enumerate all outcomes, we report the proportion of permuted data sets that have the same or smaller probabilities than that of the data. These procedures are equivalent, because the probability we calculate for each permuted data set has just that probability of arising. An empirical demonstration of the equivalence can be conducted very quickly. Five cards can be labeled AA, AA, AA, BB, and BB and then torn in half to give six A halves and four B halves. Shuffling this set of 10 cards and dealing into five pairs will produce one of only three possible data sets. The complete enumeration is so simple that permutation is not needed. However, permutation will produce the original set of all homozygotes in 1 out of 21 shuffles on average, and this is just the probability that follows from Equation 5.4.

Homozygosity Tests

The question of independence of alleles within one locus has sometimes been addressed by looking only at homozygotes. If the population has genotype proportions obeying the Hardy-Weinberg relationship, then the expected proportion of homozygotes is the sum of the squared allele proportions. A goodness-of-fit chi-square test can be conducted on two classes: homozygotes and heterozygotes. This statistic has a chi-square distribution with 1 df when the Hardy-Weinberg hypothesis is true.

The homozygosity test is not a test of the Hardy-Weinberg hypothesis. It is possible that the various homozygotes could all depart considerably from Hardy-Weinberg expectations, with some being more frequent and some being less frequent, in such a way that the departures cancel out when the sum is taken over homozygotes. We are interested in knowing whether specific genotype proportions can be estimated as products of allele proportions, and this question is not addressed by seeing whether the combined categories of "homozygote" and "heterozygote" meet expectations. Of course, if the only reasonable alternative to the null hypothesis of Hardy-Weinberg equilibrium

were one in which *all* homozygote proportions increased, then the homozygosity test would be appropriate. Such alternatives include the Wahlund effect and the loss of alleles in some homozygotes ("allelic dropout" or "null alleles"), but care would be needed to ensure that other patterns of departure from Hardy-Weinberg proportions are not possible. Care would also be needed to perform a one-tailed test–i.e., one that rejected only for *more* homozygotes than expected, rather than a two-tailed test that would reject for either more or fewer homozygotes than expected.

Multilocus Testing

There is a natural extension of the exact tests for pairs of loci. The issue becomes that of deciding whether two-locus genotype proportions can be estimated as products of allele proportions at both loci. For locus **A** with alleles A_1, A_2 and locus **B** with alleles B_1, B_2, it is convenient to use the notation of Table 4.10. There are nine genotypes, numbered 1 to 9, and the hypothesis of independence can be expressed in nine separate statements:

$$\begin{aligned}
A_1A_1B_1B_1 : P_1 &= p_{A_1}^2 p_{B_1}^2 \\
A_1A_1B_1B_2 : P_2 &= 2p_{A_1}^2 p_{B_1} p_{B_2} \\
A_1A_1B_2B_2 : P_3 &= p_{A_1}^2 p_{B_2}^2 \\
A_1A_2B_1B_1 : P_4 &= 2p_{A_1} p_{A_2} p_{B_1}^2 \\
A_1A_2B_1B_2 : P_5 &= 4p_{A_1} p_{A_2} p_{B_1} p_{B_2} \\
A_1A_2B_2B_2 : P_6 &= 2p_{A_1} p_{A_2} p_{B_2}^2 \\
A_2A_2B_1B_1 : P_7 &= p_{A_2}^2 p_{B_1}^2 \\
A_2A_2B_1B_2 : P_8 &= 2p_{A_2}^2 p_{B_1} p_{B_2} \\
A_2A_2B_2B_2 : P_9 &= p_{A_2}^2 p_{B_2}^2
\end{aligned}$$

The probability of the nine genotype counts $n_1, \ldots, n_9$ is given by the multinomial distribution. Under the assumption of independence of all alleles, the probabilities for the counts of alleles A_1, A_2 and the alleles B_1, B_2 are given by separate binomial distributions. The probability of the genotype counts conditional on the allele counts at both loci and assuming the hypothesis is true is therefore

$$\Pr(n_1, \ldots, n_9 | n_{A_1}, n_{A_2}, n_{B_1}, n_{B_2}) = \frac{n! 2^{H_A} n_{A_1}! n_{A_2}! 2^{H_B} n_{B_1}! n_{B_2}!}{(2n)!(2n) n_1! \ldots n_9!} \quad (5.8)$$

Consider the data in Table 5.3. To test for independence at loci *LDLR* and *GYPA*, the nine two-locus genotype counts are needed, and these have

been provided by Cellmark Diagnostics:

		GYPA			
		AA	*AB*	*BB*	T*otal*
LDLR	*AA*	$n_1 = 3$	$n_2 = 10$	$n_3 = 4$	17
	AB	$n_4 = 18$	$n_5 = 24$	$n_6 = 14$	56
	BB	$n_7 = 10$	$n_8 = 15$	$n_9 = 5$	30
	T*otal*	31	49	23	103

The four allele counts are

	A	*B*	Total
LDLR	90	116	206
GYPA	111	95	206

Assuming independence of all alleles, the probability of the data is

$$\frac{103!2^{56}90!116!2^{49}111!95!}{206!206!3!10!4!18!24!14!10!15!5!} = 0.00001141$$

Although this probability is obviously small, it is not clear whether the two-locus genotypic array belongs to the 5% least probable arrays under the assumption of independence. Unlike the one-locus example, there are too many possible arrays (over 50 million) to display each one with its conditional probability as we did in Table 5.7. There may even be too many to examine by computer. We use the permutation procedure, and with 10,000 permutations we found that the P-value for this data set is 0.74. The data do not lead to rejection of the null hypothesis of independence among alleles at *LDLR* and *GYPA*. Such two-locus tests are sometimes called tests for LINKAGE DISEQUILIBRIUM, but this is not quite correct because the test is affected by dependencies within and between loci and refers to sets of four alleles (two per locus). Linkage disequilibrium refers to pairs of alleles (one per locus) and is a between-locus measure. The probability in Equation 5.8 can be generalized to apply to any number of alleles and any number of loci (Box 5.8).

ESTIMATING INBREEDING COEFFICIENTS

We have remarked previously that complete independence of alleles within or between loci does not exist in nature. Although we doubt that the degree of departure in human populations is sufficient to cause applications of the product rule to be misleading, we are interested in being able to quantify

Box 5.8: General expression for the probability of genotype counts conditional on allele counts under the hypothesis of independence
For genotypes determined at a set of loci, suppose that $\{n_{la}\}$ is the set of allelic counts for the lth locus, H_l is the number of individuals heterozygous at the lth locus, and $\{n_g\}$ is the number of the gth multilocus genotype. Adding over genotypes gives the sample size:

$$\sum_g n_g = n$$

and adding over allelic counts at each locus gives twice the sample size:

$$\sum_a n_{la} = 2n$$

The conditional probability for the genotype counts, given the allelic counts and the assumption of independence, is

$$\Pr(\{n_g\}|\{n_{la}\}, l = 1, 2, \ldots) = \frac{n!}{\prod_g n_g!} \prod_l \frac{2^{H_l} \prod_a n_{la}!}{(2n)!}$$

For permutation procedures, only the part Q of this expression that varies over genotype arrays needs to be calculated. That is

$$Q = \frac{2^H}{\prod_g n_g!}$$

where $H = \sum_l H_l$ (Zaykin et al. 1995).
Although there is no limit on the number of alleles or loci over which this quantity may be calculated, in practice the numbers are limited by the sample size. J. S. Buckleton (personal communication) has noted that, for any sample size, there comes a point at which every multilocus genotype will occur either once or not at all in the sample unless the sample contains identical twins. The product of all the $n_g!$ terms is then 1, and the only term that varies over arrays is H. The test is therefore looking only at heterozygosity, and it is unlikely that it will be possible to detect departures from independence at large numbers of loci.

the extent of departure. This quantification may be more informative than the P-value for tests of independence. We use a set of quantities that have come to be known as F-STATISTICS, even though they are actually parameters. Each of these provides a measure of relationship between a pair of alleles, and is relative to some background level of relationship. Briefly, f is the extra extent to which two alleles within one individual are related when compared to pairs of alleles in different individuals but within the same subpopulation, whereas F is the extent of relatedness of alleles within an individual compared to alleles of different individuals in the whole population. The third common F-statistic, θ, measures the relationship between alleles of different individuals in one subpopulation when compared to pairs of alleles in different subpopulations.

The interpretation of matching DNA profiles needs to allow for structure within human populations. If the group of people who might reasonably be considered possible contributors to a crime stain have allele proportions very different from the rest of the population, then it could be misleading to provide analyses based on population-wide proportions. We will use the term "subpopulation" to refer to a group within a population, and assign f as the F-statistic for a subpopulation. This addresses the question of whether those two alleles are more related than pairs of alleles in different individuals in the same subpopulation. There is no requirement that this quantity has to be positive, although we conform to convention and refer to f as the WITHIN-POPULATION INBREEDING COEFFICIENT. It is affected primarily by the mating system within the subpopulation. To the extent that people tend to avoid marrying relatives, spouses are less related than random pairs of individuals, so the allele pairs within their children are less likely to be ibd (Chapter 4) than are random pairs of alleles. We might therefore expect f to be slightly negative for human populations, but we generally assume that it is very small if not zero.

Consider the contrived situation of a subpopulation founded by a set of people who were all first cousins to each other. Any two of these people are related, so their children will be inbred, according to the development in Chapter 4. If mating is at random within this subpopulation, however, there is no within-population inbreeding, and $f = 0$. Relative to other alleles in the subpopulation, the pair of alleles within one child have no special relationship. Of course when a child within the subpopulation is compared to individuals in the rest of the population there is clearly inbreeding–the alleles that child carries are much more likely to be ibd than they are for a random individual anywhere else in the population. Outside the subpopulation, there are not many people whose parents are cousins.

When the comparison is to pairs of alleles in the whole population, as opposed to a single subpopulation, we need the TOTAL INBREEDING COEFFICIENT F. It is F that we introduced as the inbreeding coefficient in Chapter 4, and we said there that ibd alleles were those that descended from the same allele in some reference population. In other words, the ibd relation was defined relative to distinct alleles in the reference population that were regarded as being not-ibd. The inbreeding coefficient for children of cousins is 1/16, relative to people whose parents are not related.

The third F-statistic is the COANCESTRY COEFFICIENT θ that we also met in Chapter 4. It refers to pairs of alleles in different individuals but in the same subpopulation, and it is relative to alleles in different individuals from different subpopulations. For the subpopulation of cousins, $\theta = 1/16$ because alleles taken from different individuals in the subpopulation have that probability of being ibd, whereas alleles from different individuals in different subpopulations are assumed to have no chance of being ibd.

Although F and θ can be regarded as probabilities of ibd, as is appropriate when genotypic and joint genotypic proportions are being formulated, other interpretations can be given to these quantities (Cockerham 1969) and it is these alternatives that may be best when it comes to estimation using data from real populations. Neither the probability nor any alternative interpretation carries any implication about the evolutionary forces that led to the numerical values of the parameters in a real population.

The notation we use is that of Cockerham (1969, 1973). An alternative notation was used by Wright (1951, 1965): F_{IS} for alleles within individuals (I) within subpopulations (S), F_{IT} for alleles within individuals (I) relative to the total (T), and F_{ST} for alleles between individuals within subpopulations (S) relative to the total (T). Although it is convenient to relate the two sets of parameters as $f = F_{IS}, F = F_{IT}, \theta = F_{ST}$, there are some slight differences. Wright defined his quantities for alleles identified by the gametes carrying them, whereas Cockerham defined his set for alleles defined by the individuals carrying them. For random mating subpopulations we can ignore the distinction.

There is a relationship among the three F-statistics:

$$f = \frac{F - \theta}{1 - \theta}$$

which illustrates that f can be zero even when there is inbreeding. This happens whenever $F = \theta$, as is the case for random-mating monoecious populations. We will be concerned with the case of large populations divided into random mating subpopulations, rather than contrived situations

like the group of cousins mentioned above. If there is some degree of reproductive isolation between the subpopulations, they may develop differences in allele proportions over time, or they may maintain some of whatever differences they had originally, and in the whole population the Wahlund effect (Chapter 4) leads to departures of genotypic proportions from products of allele proportions. This corresponds to zero f but nonzero F.

Within-population Inbreeding Coefficient

In Chapter 4 we discussed the probabilities of genotypes, and this necessarily involved the evolutionary processes that formed the population. Genetic sampling was involved, and the total inbreeding coefficient F was appropriate. If we are interested only in quantifying departures from Hardy-Weinberg in a particular subpopulation, however, we need the within-population coefficient f.

A general treatment allows for variation among fs for different genotypes. Writing the proportion of allele A_i as p_i, the proportion of A_iA_i homozygotes as P_{ii}, and the proportion of A_iA_j heterozygotes as P_{ij},

$$\left.\begin{array}{rcl} P_{ii_S} & = & p_{i_S}^2 + \sum_{j\neq i} f_{ij} p_{i_S} p_{j_S} \\ P_{ij_S} & = & 2p_{i_S}p_{j_S}(1-f_{ij}),\ \ i \neq j \end{array}\right\} \tag{5.9}$$

where the subscript S emphasizes that the equations hold for some particular subpopulation S.

For a locus with only two alleles A and B, there is only one distinct value of f: $f_{AA} = f_{AB} = f_{BB}$. For a locus with more than two alleles we could assign a common value f to all the f_{ij} values but then we find that iterative procedures are needed for maximum likelihood estimation of f (and of the p_i values)(Hill et al. 1995) An approximation to the maximum likelihood estimate is

$$\hat{f} \;=\; \frac{\sum_i (\hat{P}_{ii_S} - \hat{p}_{i_S}^2) + \frac{1}{2n}(1 - \sum_i \hat{P}_{ii_S})}{(1 - \sum_i \hat{p}_{i_S}^2) - \frac{1}{2n}(1 - \sum_i \hat{P}_{ii_S})}$$

where n is the number of individuals in the sample.

For the *LDLR* data in Tables 5.3 and 5.4

$$\begin{aligned} \hat{f} &= \frac{(\hat{P}_{AA} + \hat{P}_{BB} - \hat{p}_A^2 - \hat{p}_B^2) + \frac{1}{2n}(1 - \hat{P}_{AA} - \hat{P}_{BB})}{(1 - \hat{p}_A^2 - \hat{p}_B^2) - \frac{1}{2n}(1 - \hat{P}_{AA} - \hat{P}_{BB})} \\ &= \frac{(0.165 + 0.291 - 0.437^2 - 0.563^2) + \frac{1}{206}(1 - 0.165 - 0.291)}{(1 - 0.437^2 - 0.563^2) - \frac{1}{206}(1 - 0.165 - 0.291)} \\ &= -0.101 \end{aligned}$$

which is numerically larger than is usually found in human populations.

If the sample size correction terms involving $1/2n$ in the numerator and denominator of $\hat{f}$ are omitted, and the estimate is written as $\hat{f}$,

$$\hat{f} = \frac{(\hat{P}_{AA} + \hat{P}_{BB} - \hat{p}_A^2 - \hat{p}_B^2)}{(1 - \hat{p}_A^2 - \hat{p}_B^2)}$$

This gives us an alternative way of calculating the chi-square goodness-of-fit test statistic for testing the Hardy-Weinberg hypothesis:

$$X^2 = n\hat{f}^2$$

For the *LDLR* data, $\hat{f} = -0.105$, and $X^2 = 1.14$ as before.

Total Inbreeding Coefficient

As stated above, in the whole population, genotypic and allelic proportions are related as

$$\left.\begin{array}{rcl} P_{ii} & = & p_i^2 + Fp_i(1-p_i) \\ P_{ij} & = & 2p_ip_j(1-F), \ \ i \neq j \end{array}\right\} \tag{5.10}$$

where F is the total inbreeding coefficient. These are the relations that result when averages are taken over all replicates of the same set of evolutionary conditions.

Estimation of F cannot be accomplished simply by comparing genotypic and allele proportions in one (sub)population because of the requirement that F is a relative measure. It measures the relationship between pairs of alleles within individuals in a subpopulation relative to that for pairs of alleles in the whole population. Therefore information is needed about proportions in the subpopulations, and these data can be combined to provide information from the whole population. If data are available only from the whole population, then it is f that can be estimated, although this f is the within-population inbreeding coefficient for the whole population instead of a subpopulation, and it provides an estimate of $(F - \theta)/(1 - \theta)$. If F is wanted for the whole population, then information would be needed from other whole populations to provide the necessary basis for comparison.

Coancestry Coefficient

The coancestry coefficient θ refers to pairs of alleles in different individuals in the same subpopulation, relative to pairs of alleles in the whole population. Once again, estimation requires data from more than one subpopulation.

Otherwise there is no basis for comparison. There would also be no knowledge of the variation in allele proportions among populations.

If there is random mating within subpopulations, two alleles have the same relationship whether they are in the same or different individuals, $F = \theta$, and $f = 0$. In this case, one method of estimation is to compare allelic variation within and between populations (Weir and Cockerham 1984). Two MEAN SQUARES are calculated: MSA among subpopulations and MSW within subpopulations. For allele A and samples of size $2n$ alleles (n genotypes) from each of r subpopulations,

$$\begin{aligned} \mathrm{M}SA &= \frac{2n}{r-1}\sum_{S=1}^{r}(\hat{p}_{A_S}-\bar{p}_A)^2, \;\; r>1 \\ \mathrm{M}SW &= \frac{2n}{r(2n-1)}\sum_{S=1}^{r}\hat{p}_{A_S}(1-\hat{p}_{A_S}) \end{aligned}$$

where $\hat{p}_{A_S}$ is the sample proportion for the Sth subpopulation and

$$\bar{p}_A = \frac{1}{r}\sum_{S=1}^{r}\hat{p}_{A_S}$$

is the average proportion of A over all the subpopulations sampled.

We can show that the quantity

$$\hat{\beta} = \frac{\mathrm{M}SA-\mathrm{M}SW}{\mathrm{M}SA+(2n-1)\mathrm{M}SW}$$

provides an estimate of

$$\beta = \frac{\theta_w-\theta_a}{1-\theta_a}$$

where θ_w is the average of the θ values that apply to each of the r subpopulations, and θ_a is the average of the θ values that apply to the $r(r-1)$ pairs of subpopulations. This estimate does not depend on the unknown expected allele proportions. It will be small for data from a collection of subpopulations within one population of a single racial group because then alleles will have similar relationships within and between subpopulations. For data from subpopulations from racial groups that have been distinct for a longer evolutionary time, alleles will be much more related within than between subpopulations and β will be larger.

It is worth stressing that numerical estimates are seldom above 0.05 even when world-wide collections of subpopulation data are used (Cavalli-Sforza et al. 1994). The practical consequences of using Equations 4.20

are therefore unlikely to be substantial. It is difficult, however, to make statements that are much more precise than this because of the difficulty in estimating powers of the allele proportion in the reference population. We have said that the probability $p_A^2 + p_A(1 - p_A)\theta$ of homozygote AA may be taken to apply to a subpopulation, and we have shown that subpopulation data allow θ to be estimated. We also know that the sample proportion $\bar{p}_A$ provides an unbiased estimate for the reference value p_A. However, the square of this value is not unbiased for p_A^2 because of the variation between replicates of the evolutionary process. Estimation of powers of the allele proportions requires data from more than one subpopulation.

SUMMARY

DNA profiles almost always consist of pairs of alleles at several loci, and probabilities of these profiles are estimated most simply as the products of the individual allele probabilities. The implied assumption of allelic independence can be tested with exact tests, although hypothesis testing is not without some conceptual problems. At single loci, the independence assumption can be avoided by expressions that allow for the effects of population structure.

Chapter 6

Parentage Testing

INTRODUCTION

When we considered transfer evidence in Chapter 2, we assumed the framework of a criminal trial in which the two propositions H_p and H_d were to be considered. The subscripts referred to "prosecution" and "defense." In this chapter we will consider parentage disputes, which usually result in civil proceedings. However, we find it convenient to keep the same subscripts. For a civil trial, the plaintiff's proposition H_p will generally be the allegation of a woman that the defendant is the father of her child. Proposition H_d is still that of the defendant, and this may simply be that he is not the father. The notation has the convenient feature that it applies when parentage disputes result in criminal trials, as in cases of rape or incest. Throughout the chapter we use the terms "mother" and "father" to mean biological parents.

For a paternity case we write M for the mother of child C, and AF for the alleged father. Their genotypes will be denoted G_M, G_C, and G_{AF}, respectively. The two propositions are

H_p: AF is the father of C.
H_d: Some other man is the father of C.

Later we will consider more complex parentage analyses, often involving members of the same family. These arise in some cases of incest and in the identification of human remains.

EVALUATION OF EVIDENCE

As in Chapter 2 we use E to summarize all the genetic evidence: the genotypes G_M, G_C, and G_{AF} of mother, child and alleged father. We use I for the

nongenetic evidence, and this could include statements made by the mother and the alleged father about their relationship. Using Bayes' theorem, our interpretation of the evidence is

$$\frac{\Pr(H_p|E,I)}{\Pr(H_d|E,I)} = \frac{\Pr(E|H_p,I)}{\Pr(E|H_d,I)} \times \frac{\Pr(H_p|I)}{\Pr(H_d|I)} \tag{6.1}$$

or

$$\text{P}osterior\ odds = \text{L}ikelihood\ ratio \times \text{P}rior\ odds$$

We direct our attention to evaluation of the likelihood ratio, LR, as we do throughout the book, but we need first to mention three terms that are used in the field of parentage testing (Walker et al. 1983). The first term is the PATERNITY INDEX (PI), which is simply another name for the likelihood ratio in Equation 6.1. In simple paternity cases, the terms LR and PI are interchangeable. The second term is PROBABILITY OF PATERNITY, meaning the posterior probability of paternity, and the third is the PROBABILITY OF EXCLUSION to which we return later in the chapter.

For the probability of paternity we observe that

$$\begin{aligned} \Pr(H_d|E,I) &= 1 - \Pr(H_p|E,I) \\ \Pr(H_d|I) &= 1 - \Pr(H_p|I) \end{aligned}$$

so that Equation 6.1 can be rewritten in terms of posterior and prior probabilities of H_p, and rearranged to give

$$\Pr(H_p|E,I) = \frac{\text{L}R \times \Pr(H_p|I)}{\text{L}R \times \Pr(H_p|I) + [1 - \Pr(H_p|I)]}$$

If the prior odds are one, meaning that the prior probability of paternity is 0.5, the posterior probability of paternity is

$$\Pr(H_p|E,I) = \frac{\text{L}R}{\text{L}R + 1}$$

and this is the quantity that is referred to as the probability of paternity. The derivation of this is due to Essen-Möller (1938). We do not advocate the use of this probability of paternity because of the implicit assumption of a prior probability of 0.5, irrespective of the nongenetic evidence. We will not discuss the issue further, but refer the reader to the excellent discussion by Kaye (1990). The assumption of 50% prior probability is difficult to defend. At the least, any presentation of probabilities of paternity should

Table 6.1: Probabilities of paternity for a range of paternity index and prior probability values.

P*rior*	PI			
probability	1	10	100	1,000
0	0	0	0	0
0.001	0.001	0.00991	0.09099	0.5002501
0.010	0.010	0.09174	0.50251	0.9099181
0.100	0.100	0.52631	0.91743	0.9910803
0.500	0.500	0.90909	0.99009	0.9990010
0.900	0.900	0.98901	0.99889	0.9998889
0.990	0.990	0.99899	0.99989	0.9999899
0.999	0.999	0.99989	0.99999	0.9999990
1	1	1	1	1

be accompanied by a table, such as Table 6.1, showing the effects of different prior probabilities.

We now return to the LR and expand it in a way analogous to that in previous chapters:

$$\begin{aligned} \mathrm{L}R &= \frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)} \\ &= \frac{\Pr(G_C, G_M, G_{AF}|H_p, I)}{\Pr(G_C, G_M, G_{AF}|H_d, I)} \end{aligned}$$

From the third law of probability

$$\mathrm{L}R = \frac{\Pr(G_C|G_M, G_{AF}, H_p, I)}{\Pr(G_C|G_M, G_{AF}, H_d, I)} \times \frac{\Pr(G_M, G_{AF}|H_p, I)}{\Pr(G_M, G_{AF}|H_d, I)} \tag{6.2}$$

Neither H_p nor H_d includes information to affect our uncertainty in relation to G_M or G_{AF}, so the second ratio is one. Then

$$\mathrm{L}R = \frac{\Pr(G_C|G_M, G_{AF}, H_p, I)}{\Pr(G_C|G_M, G_{AF}, H_d, I)} \tag{6.3}$$

For the sake of brevity, we will drop I from the probabilities for the rest of the chapter.

There will be many (G_C, G_M, G_{AF}) configurations for which the numerator of the likelihood ratio will be zero. For example, if $G_M = A_iA_j$ and $G_C = A_iA_k$ with $k \neq j$ then C must have paternal allele A_k and AF cannot

be the father of C (mutations excluded) if he has genotype $G_{AF} = A_l A_m$ and $l, m \neq k$. Such situations are called EXCLUSIONS and we discuss them again later.

For the rest of our treatment of paternity calculations we will denote by A_M and A_P the maternal and paternal alleles received by child C. For homozygous children $G_C = A_i A_i$, both maternal and paternal alleles are the same, $A_M = A_P = A_i$. For heterozygous children $A_i A_j$, $i \neq j$, however, we may have doubt as to which allele is maternal and which is paternal. Ignoring any parental information, there are two possibilities: $A_M = A_i, A_P = A_j$ or $A_M = A_j, A_P = A_i$, and we will need to add probabilities for each of the two.

We consider first the numerator of the LR in Equation 6.3. Writing the child's genotype G_C as $A_M A_P$, and abbreviating Numerator by Num.:

$$\begin{aligned} \mathrm{N}um. &= \Pr(A_M A_P | G_M, G_{AF}, H_p) \\ &= \Pr(A_M | G_M, G_{AF}, H_p) \Pr(A_P | A_M, G_M, G_{AF}, H_p) \end{aligned}$$

from the third law of probability. Because the transmission of an allele from parent to child depends on the parent's genotype and not on the genotype of other individuals, we see that

$$\begin{aligned} \Pr(A_M | G_M, G_{AF}, H_p) &= \Pr(A_M | G_M) \\ \Pr(A_P | A_M, G_M, G_{AF}, H_p) &= \Pr(A_P | G_{AF}, H_p) \end{aligned}$$

Note that the first of these two equations makes use of the fact that the transmission of the maternal allele does not depend on whether the alleged father is the child's father. We can now write the numerator of the LR as

$$\mathrm{N}um. = \Pr(A_M | G_M) \Pr(A_P | G_{AF}, H_p) \tag{6.4}$$

When the child is homozygous $A_i A_i$, calculation of Equation 6.4 is straight forward:

$$\mathrm{N}um. = \Pr(A_M = A_i | G_M) \Pr(A_P = A_i | G_{AF}, H_p) \tag{6.5}$$

However, things are more complicated when the child is heterozygous $A_i A_j$ because the child's genotype by itself does not tell us which allele is maternal and which is paternal. The information may come from the mother's genotype, but in general we need to allow for both possibilities:

$$\begin{aligned} \mathrm{N}um. = &\Pr(A_M = A_i | G_M) \Pr(A_P = A_j | G_{AF}, H_p) \\ &+ \Pr(A_M = A_j | G_M) \Pr(A_P = A_i | G_{AF}, H_p) \end{aligned} \tag{6.6}$$

Evaluation of each of the terms on the right-hand side of this equation is straight forward, and is shown in the fourth column of Table 6.2 for all possible mother, child, and alleged father triples. For example, suppose the child has genotype $G_C = A_iA_j$, the mother has genotype $G_M = A_iA_j$, and the alleged father has genotype $G_{AF} = A_iA_i$. From the rules of allelic transmission:

$$\begin{aligned}
\Pr(A_M = A_i|G_M = A_iA_j) &= 0.5 \\
\Pr(A_M = A_j|G_M = A_iA_j) &= 0.5 \\
\Pr(A_P = A_i|G_{AF} = A_iA_i) &= 1 \\
\Pr(A_P = A_j|G_{AF} = A_iA_i) &= 0
\end{aligned}$$

and the numerator is 0.5 as shown in the table. The reader may find it helpful to verify the other entries in the table.

We now turn to the denominator of the LR in Equation 6.3. From the same kind of argument as before, and using the abbreviation Den. we have

$$\mathrm{D}en. \quad = \quad \Pr(A_M|G_M, H_d)\Pr(A_P|A_M, G_M, G_{AF}, H_d)$$

and it is still true that the maternal allele probability depends on the genotype only of the mother

$$\Pr(A_M|G_M, H_d) \quad = \quad \Pr(A_M|G_M)$$

Mother, Alleged Father, and Father Unrelated

The probability for the paternal allele under H_d is not straightforward when we have to take into account relationships between the father and the alleged father–either because they belong to the same family or because they belong to the same population. We take up these complications later, but for now suppose that the father is not related to either the mother or the alleged father. In this case

$$\Pr(A_P|A_M, G_M, G_{AF}, H_d) \quad = \quad \Pr(A_P|H_d)$$

and the LR denominator is

$$\mathrm{D}en. \quad = \quad \Pr(A_M|G_M)\Pr(A_P|H_d)$$

There are two cases to consider, depending on the child's genotype, as for the numerator of the LR. When the child is homozygous, $G_C = A_iA_i$, the denominator is

$$\mathrm{D}en. \quad = \quad \Pr(A_M = A_i|G_M)\Pr(A_P = A_i|H_d) \tag{6.7}$$

Table 6.2: Probabilities of child's genotype under H_p and H_d, and LR values for H_p versus H_d.

G_C	G_M	G_{AF}	$Num.$[1]	$Den.$[2]	LR	LR if $p_i = p_j = 0.1$
A_iA_i	A_iA_i	A_iA_i	1	p_i	$\frac{1}{p_i}$	10
		$A_iA_j,\ j \neq i$	$\frac{1}{2}$	p_i	$\frac{1}{2p_i}$	5
		$A_jA_k,\ k \neq i,j$	0	p_i	0	0
	A_iA_j $i \neq j$	A_iA_i	$\frac{1}{2}$	$\frac{p_i}{2}$	$\frac{1}{p_i}$	10
		$A_iA_j,\ j \neq i$	$\frac{1}{4}$	$\frac{p_i}{2}$	$\frac{1}{2p_i}$	5
		$A_jA_k,\ k \neq i,j$	0	$\frac{p_i}{2}$	0	0
A_iA_j $i \neq j$	A_iA_i	A_jA_j	1	p_j	$\frac{1}{p_j}$	10
		$A_jA_k,\ k \neq j$	$\frac{1}{2}$	p_j	$\frac{1}{2p_j}$	5
		$A_kA_l,\ k,l \neq j$	0	p_j	0	0
	A_iA_j $i \neq j$	A_iA_i	$\frac{1}{2}$	$\frac{p_i+p_j}{2}$	$\frac{1}{p_i+p_j}$	5
		A_iA_j	$\frac{1}{2}$	$\frac{p_i+p_j}{2}$	$\frac{1}{p_i+p_j}$	5
		$A_jA_k,\ k \neq i,j$	$\frac{1}{4}$	$\frac{p_i+p_j}{2}$	$\frac{1}{2(p_i+p_j)}$	2.5
		$A_kA_l,\ k,l \neq i,j$	0	$\frac{p_i+p_j}{2}$	0	0
	A_iA_k $k \neq i,j$	A_jA_j	$\frac{1}{2}$	$\frac{p_j}{2}$	$\frac{1}{p_j}$	10
		$A_jA_l,\ l \neq j$	$\frac{1}{4}$	$\frac{p_j}{2}$	$\frac{1}{2p_j}$	5
		$A_kA_l,\ k,l \neq j$	0	$\frac{p_j}{2}$	0	0

[1] $Num. = \Pr(G_C|G_M, G_{AF}, H_p)$
[2] $Den. = \Pr(G_C|G_M, G_{AF}, H_d)$

and when the child is heterozygous, $G_C = A_iA_j$,

$$\begin{aligned} \text{D}en. &= \Pr(A_M = A_i|G_M)\Pr(A_P = A_j|H_d) \\ &\quad + \Pr(A_M = A_j|G_M)\Pr(A_P = A_i|H_d) \end{aligned} \tag{6.8}$$

The numerical values for these two equations are shown in the fifth column of Table 6.2, and we consider one case for illustration. Suppose the child has genotype $G_C = A_iA_j$ and the mother has genotype $G_M = A_iA_j$. For Equation 6.8 we use

$$\begin{aligned} \Pr(A_M = A_i|G_M = A_iA_j) &= 0.5 \\ \Pr(A_M = A_j|G_M = A_iA_j) &= 0.5 \end{aligned}$$

We also use

$$\begin{aligned} \Pr(A_P = A_i|H_d) &= p_i \\ \Pr(A_P = A_j|H_d) &= p_j \end{aligned}$$

and this leads to value of $(p_i + p_j)/2$ for the denominator of the LR, as given in the table.

The LR values for all combinations of genotypes of mother, child, and alleged father are shown in Table 6.2, along with some illustrative values for allele frequencies $p_i = p_j$ of 0.1.

Hypotheses Specifying Relation of Alleged Father

In the previous section we assumed that the alternative hypothesis was

> H_d: Some unknown man, unrelated to the alleged father, was the father.

Now we consider the case where

> H_d: A relative of the alleged father is the father.

We will continue to assume that the mother and alleged father are unrelated. The calculations can make use of four-allele descent measures we introduced in Chapter 4. The problem is often simpler here, however, because there are just three alleles to consider: one from the father and the two in the alleged father. Specifically, we need to evaluate for the denominator:

$$\begin{aligned} \Pr(A_P|A_M, G_M, G_{AF}, H_d) &= \Pr(A_P|G_{AF}, H_d) \\ &= \frac{\Pr(A_P, G_{AF})}{\Pr(G_{AF})} \end{aligned}$$

Table 6.3: Probabilities of allele triples in terms of allele frequencies and descent measures.

$c = A_P$	$ab = G_{AF}$	γ_{abc}	$\gamma_{bc} + \gamma_{ac}$	γ_{ab}	γ_0
A_i	A_iA_i	p_i	p_i^2	p_i^2	p_i^3
A_i	$A_jA_j,\ j \neq i$	0	0	p_ip_j	$p_ip_j^2$
A_i	$A_iA_j,\ j \neq i$	0	p_ip_j	0	$2p_i^2p_j$
A_i	$A_jA_k,\ j,k \neq i$	0	0	0	$2p_ip_jp_k$

We have dropped the dependencies on A_M, and G_M because of the assumption that the child's parents are unrelated. The numerator of the right-hand side requires information about a set of three alleles, and the denominator is the proportion of genotype G_{AF} in the population.

For three alleles, a, b, and c, there are five patterns of identity by descent. We retain the equivalence sign to indicate identity by descent (ibd) and write the five probabilities as

$$\begin{aligned}
\gamma_{abc} &= \Pr(a \equiv b \equiv c) \\
\gamma_{bc} &= \Pr(only\ b \equiv c) \\
\gamma_{ac} &= \Pr(only\ a \equiv c) \\
\gamma_{ab} &= \Pr(only\ a \equiv b) \\
\gamma_0 &= \Pr(none\ ibd)
\end{aligned}$$

The three alleles referred to by these five measures will now be identified by the individuals that carry them, namely the alleged father and the father. For the case when a, b are the two alleles carried by the alleged father AF and c is the paternal allele (i.e., one of the two alleles carried by the father TF) we find it useful to write

$$\begin{aligned}
\gamma_{abc} &= \gamma_{\ddot{A}T} \\
\frac{1}{2}(\gamma_{ac} + \gamma_{bc}) &= \theta_{AT} - \gamma_{\ddot{A}T} \\
\gamma_{ab} &= F_A - \gamma_{\ddot{A}T} \\
\gamma_0 &= 1 - 2\theta_{AT} - F_A + 2\gamma_{\ddot{A}T}
\end{aligned}$$

where we have abbreviated AF by A and TF by T. The quantity $\gamma_{\ddot{A}T}$ is the probability of both alleles in AF and one allele in TF are all ibd. We met the coancestry θ_{AT} of AF and TF and the inbreeding coefficient F_A of AF in Chapter 4. It is usually sufficient to work with the average of $\gamma_{ac} + \gamma_{bc}$, instead of keeping them separated.

The three-allele descent measures serve to provide probabilities of sets of specific types of alleles, just as the two-allele measures gave genotype proportions and the four-allele measures gave joint genotype proportions. Suppose that both men AF and TF belong to a population in which alleles of type A_i have proportion p_i. Then, if a, b are the alleles in AF, $G_{AF} = ab$, and $c = A_P$ is a random allele in TF, the values of $\Pr(a, b, c)$ require an allele probability for each distinct allele or set of ibd alleles as shown in Table 6.3. From the first row in that table,

$$\Pr(a = A_i, bc = A_iA_i) \quad = \quad \gamma_{abc}p_i + (\gamma_{bc} + \gamma_{ac} + \gamma_{ab})p_i^2 + \gamma_0 p_i^3$$

Using the $\gamma_{\ddot{A}T}, \theta_{AT}, F_A$ formulation, and recalling that the probability of AF being homozygous A_iA_i is $p_i^2 + F_A p_i(1 - p_i)$, it can be shown that

$$\begin{aligned}\Pr(A_P = A_i | G_{AF} = A_iA_i, H_d) \quad &= \quad \frac{p_i^2 + (2\theta_{AT} + F_A)p_i(1 - p_i)}{p_i + F_A(1 - p_i)} \\ &\quad + \frac{\gamma_{\ddot{A}T}(1 - p_i)(1 - 2p_i)}{p_i + F_A(1 - p_i)}\end{aligned}$$

The probabilities for other combinations of paternal allele and alleged father genotype are

$$\begin{aligned}\Pr(A_P = A_i | G_{AF} = A_jA_j, H_d) \quad &= \quad \frac{p_i(F_A - \gamma_{\ddot{A}T}) + p_i p_j X}{p_j + F_A(1 - p_j)} \\ \Pr(A_P = A_i | G_{AF} = A_iA_j, H_d) \quad &= \quad \frac{(\theta_{AT} - \gamma_{\ddot{A}T}) + p_i X}{1 - F_A} \\ \Pr(A_P = A_i | G_{AF} = A_jA_k, H_d) \quad &= \quad \frac{p_i X}{1 - F_A}\end{aligned}$$

where

$$X \quad = \quad (1 - 2\theta_{AT} - F_A + 2\gamma_{\ddot{A}T})$$

In each case, different subscripts indicate different alleles. The usual situation is that the alleged father is not inbred, so that $F_A = \gamma_{\ddot{A}T} = 0$, and

then

$$\begin{aligned}
\Pr(A_P = A_i | G_{AF} = A_i A_i) &= p_i(1 - 2\theta_{AT}) + 2\theta_{AT} \\
\Pr(A_P = A_i | G_{AF} = A_j A_j) &= p_i(1 - 2\theta_{AT}),\ i \neq j \\
\Pr(A_P = A_i | G_{AF} = A_i A_j) &= p_i(1 - 2\theta_{AT}) + \theta_{AT},\ i \neq j \\
\Pr(A_P = A_i | G_{AF} = A_j A_k) &= p_i(1 - 2\theta_{AT}),\ i \neq j, k
\end{aligned}$$

The likelihood ratios for H_p versus H_d are shown in Table 6.4 for the situation without inbreeding. The numerical values are for $\theta_{AT} = 0.25$, corresponding to the case where the alleged father is a brother to the father.

Exercise 6.1 The *Gc* system has produced the following three genotypes in a case of disputed paternity: $G_M = AC, G_C = AB$, and $G_{AF} = BC$. Using the sample allele proportions in Table 5.4, calculate the LR in the cases where (a) AF makes no allegation about the identity of the true father; (b) AF alleges that the true father is his first cousin.

Exercise 6.2 Repeat Exercise 6.1 for the case where $G_M = AB, G_C = AB$, and $G_{AF} = BC$. For part (b), suppose that AF alleges his half brother is the true father.

Avuncular Index

Morris et al. (1988) considered the situation that arises when the alleged father cannot be typed, but typing is available for his relative R. The alternative pair of propositions now involve this relative rather than the alleged father:

H_p: The father of C is a relative of R.
H_d: The father of C is unrelated to R.

Another way of expressing H_p in this case is to specify the relationship between R and the child. For example, if R's brother is the child's father, then R is the child's uncle and Morris et al. coined the phrase "avuncular index" as an alternative to paternity index. We will continue to use the term likelihood ratio, however, as it applies in all cases.

The likelihood ratio is

$$\mathrm{L}R = \frac{\Pr(G_C | G_M, G_R, H_p)}{\Pr(G_C | G_M, G_R, H_d)}$$

Table 6.4: LR values when there is a relationship, measured by θ_{AT}, between the father and the alleged father under H_d.

G_C	G_M	G_{AF}	LR	LR *if* $\theta_{AT} = 0.25$ *and* $p_i = p_j = 0.1$
A_iA_i	A_iA_i	A_iA_i	$\frac{1}{p_i(1-2\theta_{AT})+2\theta_{AT}}$	1.82
		$A_iA_j,\ j \neq i$	$\frac{1}{2[p_i(1-2\theta_{AT})+\theta_{AT}]}$	1.67
		$A_jA_k,\ k \neq i,j$	0	0
	A_iA_j $i \neq j$	A_iA_i	$\frac{1}{p_i(1-2\theta_{AT})+2\theta_{AT}}$	1.82
		$A_iA_j,\ j \neq i$	$\frac{1}{2[p_i(1-2\theta_{AT})+\theta_{AT}]}$	1.67
		$A_jA_k,\ k \neq i,j$	0	0
A_iA_j $i \neq j$	A_iA_i	A_jA_j	$\frac{1}{p_j(1-2\theta_{AT})+2\theta_{AT}}$	1.82
		$A_jA_k,\ k \neq j$	$\frac{1}{2[p_j(1-2\theta_{AT})+\theta_{AT}]}$	1.67
		$A_kA_l,\ k,l \neq j$	0	0
	A_iA_j $i \neq j$	A_iA_i	$\frac{1}{(p_i+p_j)(1-2\theta_{AT})+2\theta_{AT}}$	1.67
		A_iA_j	$\frac{1}{(p_i+p_j)(1-2\theta_{AT})+2\theta_{AT}}$	1.67
		$A_jA_k,\ k \neq i,j$	$\frac{1}{2(p_i+p_j)(1-2\theta_{AT})+2\theta_{AT}}$	1.43
		$A_kA_l,\ k,l \neq i,j$	0	0
	A_iA_k $k \neq i,j$	A_jA_j	$\frac{1}{p_j(1-2\theta_{AT})+2\theta_{AT}}$	1.82
		$A_jA_l,\ l \neq j$	$\frac{1}{2[p_j(1-2\theta_{AT})+\theta_{AT}]}$	1.67
		$A_kA_l,\ k,l \neq j$	0	0

$$= \frac{\Pr(A_M|G_M)\Pr(A_P|G_R,H_p)}{\Pr(A_M|G_M)\Pr(A_P|H_d)}$$

which is obviously related to the likelihood ratios in Table 6.4. If those values were written as PI, for paternity index, and the present values written as AI for avuncular index, then Morris et al. (1988) noticed that

$$\mathrm{A}I = (1-2\theta_{AR}) + 2\theta_{AR}\mathrm{P}I$$

where θ_{AR} is the coancestry coefficient for the alleged father and the tested relative. This relationship requires the absence of inbreeding. Now that the tested man is not being alleged to be the father, he need not carry the paternal allele, so it may be that PI is zero but AI is not zero. Recall that θ_{AR} takes the value 1/4 for brothers or for father and son, 1/8 for half-brothers or for uncle and nephew, and 1/16 for first cousins.

Incestuous Paternity

When the mother and alleged father are closely related, the paternity issue becomes a criminal matter. The particular relationships that are prohibited by law generally correspond to coancestry coefficients greater than 1/16. If the alternative proposition H_d is that the father is an unknown man unrelated to both mother and alleged father, then the analysis proceeds as in Table 6.2. The relationship of AF and M does not affect the probability of the child's genotype under H_p or H_d. However, it is not uncommon in such situations for H_d to be that the father is some other relative of the mother.

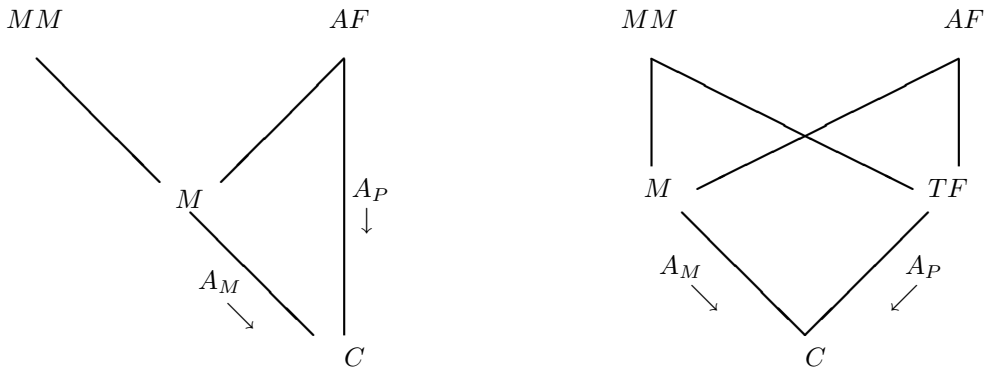
Alleged father is mother's father. Suppose the alleged father AF is also the mother's father:

H_p: AF, the father of M, is the father of C.

If the alternative proposition H_d is that the father of C is an unknown man, unrelated to both M and AF, then the likelihood ratio is just as shown in Table 6.2. However, if H_d is

H_d: TF, another relative of M, is the father of C.

then the analysis is more complex. We will consider the case where TF is a brother of M and a son of AF, as shown in Figure 6.1, *and he has not been tested*.



Proposition H_p Proposition H_d

Figure 6.1: Alleged father is mother's father.

As in Equation 6.2, the likelihood ratio is

$$\mathrm{L}R \;=\; \frac{\Pr(G_C|G_M, G_{AF}, H_p)}{\Pr(G_C|G_M, G_{AF}, H_d)} \times \frac{\Pr(G_M, G_{AF}|H_p)}{\Pr(G_M, G_{AF}|H_d)}$$

The second ratio on the right hand side cancels to one, and

$$\mathrm{L}R \;=\; \frac{\Pr(G_C|G_M, G_{AF}, H_p)}{\Pr(G_C|G_M, G_{AF}, H_d)}$$

Evaluation of the numerator proceeds as for the simple paternity case, and the values for homozygous or heterozygous children are given by Equations 6.5 or 6.6.

Calculation of the LR denominator is more complicated. If $G_C = A_M A_P$, we have

$$\mathrm{D}en. \;=\; \Pr(A_M|G_M, G_{AF}, H_d)\Pr(A_P|A_M, G_M, G_{AF}, H_d)$$

and the first term on the right-hand side is still $\Pr(A_M|G_M)$ because A_M is known to have come from the mother with genotype G_M. Also, the paternal allele is independent of the maternal allele so

$$\mathrm{D}en. \;=\; \Pr(A_M|G_M)\Pr(A_P|G_M, G_{AF}, H_d)$$

Remember that if there is doubt as to which of the child's alleles is maternal and which is paternal, we need to sum over both possibilities.

To evaluate $\Pr(A_P|G_M, G_{AF}, H_d)$ we need to introduce the genotype G_{MM} of the mother's mother *MM*. In effect, this allows statements to be made about the genotype of the mother's brother *TF*. Unless *MM* has been typed, her genotype is not known and we must use the law of total probability to sum over all possible values of G_{MM}. For simplicity we drop H_d from the notation, and write

$$\Pr(A_P|G_M, G_{AF}) = \sum_{G_{MM}} \Pr(A_P|G_{MM}, G_M, G_{AF}) \Pr(G_{MM}|G_M, G_{AF})$$

The paternal allele does not depend on the mother's genotype, so G_M can be removed from the conditioning in the first term of the right-hand side. The LR denominator becomes

$$\mathrm{D}en. = \Pr(A_M|G_M) \sum_{G_{MM}} \Pr(A_P|G_{MM}, G_{AF}) \Pr(G_{MM}|G_M, G_{AF})$$

We can evaluate this because, once G_M and G_{AF} are known, we can calculate the probabilities of all possible values of G_{MM}. Until now we have traced the transmission of alleles forward in time, from parent to child. Now, to make inferences about the genotype of *MM* we have to move backward in time from the genotype of her child *M*. Use of Bayes' theorem allows us to express this backward-moving probability in terms of the forward-moving probabilities that we know from genetic laws:

$$\begin{aligned} \Pr(G_{MM}|G_M, G_{AF}) &= \frac{\Pr(G_M|G_{MM}, G_{AF}) \Pr(G_{MM}|G_{AF})}{\Pr(G_M|G_{AF})} \\ &= \frac{\Pr(G_M|G_{MM}, G_{AF}) \Pr(G_{MM})}{\Pr(G_M|G_{AF})} \end{aligned}$$

We have replaced $\Pr(G_{MM}|G_{AF})$ by $\Pr(G_{MM})$ in the numerator of the right-hand side because the parents of *M* are assumed to be unrelated. Of course, if *MM* has been typed, the calculations are just the usual transmission probabilities and there is no need to sum over the unknown genotypes G_{MM}. Otherwise, the LR is

$$\mathrm{L}R = \frac{\Pr(A_M|G_M) \Pr(A_P|G_{AF}, H_p)}{\mathrm{D}en.}$$

where

$$\begin{aligned} \mathrm{D}en. = {} & \frac{\Pr(A_M|G_M)}{\Pr(G_M|G_{AF})} \\ & \times \sum_{G_{MM}} \Pr(A_P|G_{MM}, G_{AF}, H_d) \Pr(G_M|G_{MM}, G_{AF}) \Pr(G_{MM}) \end{aligned}$$

Table 6.5: Probabilities needed for situation in Figure 6.1.

.

G_{MM}	$\Pr(G_{MM})$	$\Pr(G_M \mid G_{MM}, G_{AF})$	$\Pr(A_P \mid G_{MM}, G_{AF})$
A_jA_j	p_j^2	1/2	1/4
A_iA_j	$2p_ip_j$	1/4	1/4
A_jA_k	$2p_jp_k$	1/4	1/2
A_jA_l $l \neq i,j,k$	$2p_jp_l$	1/4	1/4

and we resist the temptation to cancel the $\Pr(A_M|G_M)$ terms because in some cases we need to sum the numerator and denominator over two values of A_M.

To make the analysis concrete, suppose that $G_M = A_iA_j, G_C = A_iA_k$, and $G_{AF} = A_iA_k$. This is a situation where the maternal and paternal alleles can be deduced without ambiguity: $A_M = A_i, A_P = A_k$. The mother's mother must therefore carry allele $A_M = A_j$. Furthermore, $\Pr(G_M|G_{AF}) = p_j/2$. The other probabilities depend on the genotype of *MM*, as shown in Table 6.5. Therefore the likelihood ratio is

$$\mathrm{L}R \;=\; \frac{2}{1+p_k}$$

Exercise 6.3 Repeat the analysis of this section in the case where the mother, the alleged father, and the child all have the same genotype A_iA_j.

Structured Populations

For structured populations, there is a low level of relatedness between all members of the same subpopulation. We met this situation in Chapter 4 and now consider the implications for parentage testing. The mother, alleged father, and father, although not in the same family, have some relatedness by virtue of belonging to the same subpopulation. If allele proportions are known for this subpopulation we can use the results of Table 6.1, but if all we know are the allele proportions in the whole population we need to take account of the genetic variability among subpopulations.

There is no change to $\Pr(G_C|G_M, G_{AF}, H_p)$ since the genotypes of M and AF give the probability of the genotype of C. Under H_d we can no longer assume that the maternal and paternal alleles are independent. If the paternal allele A_P is from man TF,

$$\begin{aligned}\Pr(G_C|G_M, G_{AF}, H_d) &= \Pr(A_M A_P|G_M, G_{AF}, H_d) \\ &= \Pr(A_M|G_M, G_{AF}, H_d)\Pr(A_P|A_M, G_M, G_{AF}, H_d) \\ &= \Pr(A_M|G_M)\Pr(A_P|G_M, G_{AF}, H_d)\end{aligned}$$

Therefore we need the probability of A_P being an allele in an unknown man TF in the population given the genotypes of the mother and the alleged father. The quantity $\Pr(A_M|G_M)$ is just 0, 1/2, or 1 as before. The conditional probability needed is

$$\Pr(A_P|G_M, G_{AF}, H_d) = \frac{\Pr(A_P, G_M, G_{AF})}{\Pr(G_M, G_{AF})}$$

and the numerator of the right-hand side requires the relationships among sets of five alleles: the paternal allele and the two alleles in each of the mother and the alleged father. The denominator requires the relationship among sets of four alleles: the two in each of the mother and the alleged father. These probabilities are generally difficult to determine, but simple expressions hold for random-mating populations that have reached an evolutionary equilibrium as shown in the section on arbitrary sets of alleles in Chapter 4.

The likelihood ratio for

H_p: AF is the father of C.
H_d: AF is not related to C.

depends on the genotypes of the mother and alleged father and the paternal allele type, rather than just on the paternal allele and the genotype of the alleged father. The results are shown in Table 6.6 (Balding and Nichols 1995).

As in previous sections, we need an additional step when the maternal and paternal alleles are not determined uniquely. In general,

$$\begin{aligned}\mathrm{L}R &= \frac{\Pr(G_C|G_M, G_{AF}, H_p)}{\Pr(G_C|G_M, G_{AF}, H_d)} \\ &= \frac{\sum_{A_M,A_P}\Pr(A_M|G_M, G_{AF}, H_p)\Pr(A_P|G_M, G_{AF}, H_p)}{\sum_{A_M,A_P}\Pr(A_M|G_M, G_{AF}, H_d)\Pr(A_P|G_M, G_{AF}, H_d)} \\ &= \frac{\sum_{A_M,A_P}\Pr(A_M|G_M)\Pr(A_P|G_{AF}, H_p)}{\sum_{A_M,A_P}\Pr(A_M|G_M)\Pr(A_P|G_M, G_{AF}, H_d)}\end{aligned}$$

Table 6.6: LR values for the alternative propositions that the alleged father either is or is not the father when mother, father, and alleged father all belong to the same subpopulation. (Different subscripts indicate different alleles.) The proportions p_i refer to the whole population.

G_M	G_C	A_M	A_P	G_{AF}	PI	PI *when* $\theta = 0.03, p_i = 0.1$
A_iA_i	A_iA_i	A_i	A_i	A_iA_i	$\frac{1+3\theta}{4\theta+(1-\theta)p_i}$	5.0
				A_iA_j	$\frac{1+3\theta}{2[3\theta+(1-\theta)p_i]}$	3.0
	A_iA_j	A_i	A_j	A_jA_j	$\frac{1+3\theta}{2\theta+(1-\theta)p_j}$	6.6
				A_iA_j	$\frac{1+3\theta}{2[\theta+(1-\theta)p_j]}$	4.5
				A_jA_k	$\frac{1+3\theta}{2[\theta+(1-\theta)p_j]}$	4.5
A_iA_k	A_iA_i	A_i	A_i	A_iA_i	$\frac{1+3\theta}{3\theta+(1-\theta)p_i}$	6.0
				A_iA_k	$\frac{1+3\theta}{2[2\theta+(1-\theta)p_i]}$	3.6
	A_iA_j	A_i	A_j	A_jA_j	$\frac{1+3\theta}{2\theta+(1-\theta)p_j}$	6.6
				A_iA_j	$\frac{1+3\theta}{2[\theta+(1-\theta)p_j]}$	4.5
				A_jA_l	$\frac{1+3\theta}{2[\theta+(1-\theta)p_j]}$	4.5

Now suppose that mother, child, and alleged father all have the same genotype A_iA_j. The numerator of LR is 0.5, and Equation 4.23 provides a value for the denominator of

$$\begin{aligned}&\Pr(A_M = A_i|G_M = A_iA_j)\Pr(A_P = A_j|G_M = G_{AF} = A_iA_j, H_d)\\&+\Pr(A_M = A_j|G_M = A_iA_j)\Pr(A_P = A_i|G_M = G_{AF} = A_iA_j, H_d)\end{aligned}$$

and this has the value

$$\frac{1}{2}\left[\frac{2\theta + (1-\theta)p_i}{1+3\theta} + \frac{2\theta + (1-\theta)p_j}{1+3\theta}\right]$$

The LR becomes

$$\mathrm{L}R \;=\; \frac{1+3\theta}{4\theta + (1-\theta)(p_i + p_j)}$$

PATERNITY EXCLUSION

In paternity disputes the question is whether or not a particular man is the father of a particular child. Classical considerations of such questions were limited to excluding a man from paternity of a child when the man did not have the child's paternal allele at some locus, or, if the paternal allele cannot be determined, when the man had neither of the child's alleles. The increasing availability of diagnostic loci has given rise to calculations based on the probabilities of the child's genotype under alternative propositions as we have shown. For completeness, however, we now review some results pertaining to exclusion.

In Table 6.7 we show all the possible combinations of genotypes of a mother and her child for a locus with alleles A_i. The last column of the table shows the probability that an unknown man will be excluded as being the father of the child. This calculation has nothing to do with any specific alleged father, and it is seen to depend on the population proportion of the allele inferred to be the child's paternal allele. Except for the case where both mother and child have the same heterozygous genotype, the paternal allele is easily identified as the allele the child did not get from the mother. For the double heterozygote case, either allele could have been the paternal allele, so the exclusion probability uses the sum of the proportions of both alleles. If there are only two alleles, no man could be excluded from paternity by this locus.

A genetic marker can be characterized by its ability to exclude an unrelated man from paternity in any situation. This exclusion probability is

Table 6.7: Paternity exclusion configurations at one locus with an arbitrary number of alleles; k is any value different from i and j.

Mother		*Child*		*Excluded man*	
Type	*Probability*	*Type*	*Probability*[1]	*Genotypes*	*Probability*
A_iA_i	p_i^2	A_iA_i	p_i	$A_wA_x,\ w,x \neq i$	$(1-p_i)^2$
		A_iA_j	p_j	$A_wA_x,\ w,x \neq j$	$(1-p_j)^2$
A_iA_j	$2p_ip_j$	A_iA_i	$p_i/2$	$A_wA_x,\ w,x \neq i$	$(1-p_i)^2$
$j \neq i$		A_jA_j	$p_j/2$	$A_wA_x,\ w,x \neq j$	$(1-p_j)^2$
		A_iA_j	$(p_i+p_j)/2$	$A_wA_x,\ w,x \neq i,j$	$(1-p_i-p_j)^2$
		A_iA_k	$p_k/2$	$A_wA_x,\ w,x \neq k$	$(1-p_k)^2$
		A_jA_k	$p_k/2$	$A_wA_x,\ w,x \neq k$	$(1-p_k)^2$

[1]*Probability of genotype of child given genotype of mother.*

given by summing the joint probabilities of all the mother-child-excluded man combinations shown in the table. The probability of an A_iA_i mother with an A_iA_i child is $p_i^2 \times p_i$, and this combination excludes all men that do not have an A_i allele. Such men occur in proportion $(1-p_i)^2$, so that the trio in the first line of Table 6.7 has a combined probability of $p_i^3(1-p_i)^2$. Adding such probabilities from all seven lines of the table gives the exclusion probability Q

$$Q = \sum_i p_i(1-p_i)^2 - \frac{1}{2}\sum_i \sum_{j \neq i} p_i^2 p_j^2 (4 - 3p_i - 3p_j)$$

This probability will be maximized when all m alleles at the locus have proportion $1/m$.

$$Q_{max} = 1 - \frac{2m^3 + m^2 - 5m + 3}{m^4}$$

These exclusion probabilities depend on allele frequencies for that locus, but do not depend on the genotypes in any particular case. The utility of a locus increases with the number of alleles, although even with 5 alleles the value of Q_{max} is only 0.6.

Exclusion probabilities are increased with the use of several loci, since it is sufficient to exclude at any one of several loci when mutation can be

ignored. If Q_l is the exclusion probability at locus l, then the overall probability of exclusion follows from being able to exclude from at least one locus. In other words, Q is one minus the probability that none of the loci allows exclusion. If the Q_l are independent

$$Q = 1 - \prod_l (1 - Q_l)$$

For two loci, each with five equally frequent alleles, the value of Q increases to $[1 - (1 - 0.6)^2] = 0.84$, and for five such loci it is 0.99.

MISSING PERSONS

Calculations similar to those used in paternity disputes are made when DNA is recovered from stains or remains thought to be from a missing person. In a forensic setting, a bloodstain may be found in the car of a person suspected of having abducted and disposed of a victim. After an aircraft crash, body parts may be found scattered around the wreckage. In a military setting, bones may be returned from a foreign country many years after a war was fought in that country. In each case, the profile from the sample is compared to profiles from people known to be immediate family members of the missing person. Generally the propositions to be considered are that the sample is from the missing person or that it is from some unrelated person.

Spouse and Child Typed

Suppose a person is missing. The genetic evidence E consists of the genotype of the sample from X, who may be the missing person, together with the genotypes from the spouse M and child C of the missing person. The two alternative propositions are:

H_p: The sample is from the missing person.
H_d: The sample is from some unknown person.

The likelihood ratio is

$$\begin{aligned} \mathrm{L}R &= \frac{\Pr(E|H_p)}{\Pr(E|H_d)} \\ &= \frac{\Pr(G_C, G_M, G_X|H_p)}{\Pr(G_C, G_M, G_X|H_d)} \end{aligned}$$

Table 6.8: Probabilities needed in missing person case.

Child G_C	S*pouse* G_M	S*ample* G_X	$\Pr(G_C \mid G_M, G_X, H_p)$	$\Pr(G_C \mid G_M, H_d)$
A_iA_i	A_iA_i	A_iA_i	1	p_i
		A_iA_j	0.5	p_i
	A_iA_j	A_iA_k	0.25	$0.5p_i$
A_iA_j	A_iA_i	A_jA_j	1	p_j
		A_iA_j	0.5	p_j
	A_iA_k	A_jA_j	0.5	$0.5p_j$
		A_jA_k	0.25	$0.5p_j$

and we proceed by working with probabilities of genotypes conditional on those in the previous generation(s):

$$\begin{aligned} LR &= \frac{\Pr(G_C \mid G_M, G_X, H_p)\Pr(G_M, G_X \mid H_p)}{\Pr(G_C \mid G_M, G_X, H_d)\Pr(G_M, G_X \mid H_d)} \\ &= \frac{\Pr(G_C \mid G_M, G_X, H_p)}{\Pr(G_C \mid G_M, H_d)} \end{aligned}$$

since the genotype of the child does not depend on that of X when H_d is true. We ignore the low level of relatedness due to population structure. The probabilities for those cases when H_p is not contradicted by the genetic evidence are shown in Table 6.8, so the likelihood ratios are the same as in the paternity case in which X is alleged to be the father of child C who has mother M. As in paternity disputes, extensions can be made to allow for X to be a relative of the missing person, or to allow for relatedness among all members of a population.

Additional Family Typings

It may be the case that family members apart from the spouse and child of the missing person are typed. The general procedure is the same: the probabilities of the set of observed genotypes under two propositions are compared.

Suppose the parents P and Q as well as the child C and spouse M of the missing person are typed, and that a sample X is available. Hypothesis H_p

is that X is from the missing person. Under proposition H_d the sample is from some unknown person, and therefore the genotype G_X of X does not depend on the genotypes G_P and G_Q of P and Q, and the genotype of G_C of C does not depend on G_X.

In order to invoke the laws of genetic transmission, the likelihood ratio is arranged to involve probabilities of genotypes conditional on previous generations. The first step is to invoke the probability of the child conditional on its parent(s):

$$\begin{aligned} \mathrm{L}R &= \frac{\Pr(E|H_p)}{\Pr(E|H_d)} \\ &= \frac{\Pr(G_C, G_M, G_X, G_P, G_Q|H_p)}{\Pr(G_C, G_M, G_X, G_P, G_Q|H_d)} \\ &= \frac{\Pr(G_C|G_M, G_X, G_P, G_Q, H_p)\Pr(G_M, G_X, G_P, G_Q|H_p)}{\Pr(G_C|G_M, G_X, G_P, G_Q, H_d)\Pr(G_M, G_X, G_P, G_Q|H_d)} \end{aligned}$$

When the child's parents are known, their genotypes are sufficient to specify that of a child, so the child's grandparental genotypes are not needed. In other words

$$\Pr(G_C|G_M, G_X, G_P, G_Q, H_p) = \Pr(G_C|G_M, G_X, H_p)$$

If the sample is not from the missing person, the genotype G_X has no effect on the probability of G_C, but the genotypes G_P and G_Q are needed:

$$\Pr(G_C|G_M, G_X, G_P, G_Q, H_d) = \Pr(G_C|G_M, G_P, G_Q, H_d)$$

These last two results lead to

$$\mathrm{L}R = \frac{\Pr(G_C|G_M, G_X, H_p)\Pr(G_M, G_X, G_P, G_Q|H_p)}{\Pr(G_C|G_M, G_P, G_Q, H_d)\Pr(G_M, G_X, G_P, G_Q|H_d)}$$

The next step is to make the probabilities of the genotypes of X and M conditional on those of the parents - in line with the direction of flow of genetic information. Under H_p, when X is from the offspring of P and Q,

$$\begin{aligned} \Pr(G_M, G_X, G_P, G_Q|H_p) &= \Pr(G_M, G_X|G_P, G_Q, H_p)\Pr(G_P, G_Q|H_p) \\ &= \Pr(G_M|H_p)\Pr(G_X|G_P, G_Q, H_p) \\ &\quad \times \Pr(G_P, G_Q|H_p) \end{aligned}$$

because the spouse's genotype G_M does not depend on the genotypes of the missing person and is also independent of the genotype G_X. Under H_d,

when X is from some unknown person,

$$\begin{aligned}\Pr(G_M, G_X, G_P, G_Q|H_d) &= \Pr(G_M, G_X|G_P, G_Q, H_d)\Pr(G_P, G_Q|H_d) \\ &= \Pr(G_M|H_d)\Pr(G_X|H_d)\Pr(G_P, G_Q|H_d)\end{aligned}$$

These two results lead to

$$\begin{aligned}\mathrm{L}R &= \frac{\Pr(G_C|G_M, G_X, H_p)\Pr(G_M|H_p)\Pr(G_X|G_P, G_Q, H_p)}{\Pr(G_C|G_M, G_P, G_Q, H_d)\Pr(G_M|H_d)\Pr(G_X|H_d)} \\ &\quad \times \frac{\Pr(G_P, G_Q|H_p)}{\Pr(G_P, G_Q|H_d)}\end{aligned}$$

The final step is to recognize that the genotypes G_M, G_P, and G_Q do not depend on the propositions, so they cancel from numerator and denominator of the LR:

$$\mathrm{L}R = \frac{\Pr(G_C|G_M, G_X, H_p)\Pr(G_X|G_P, G_Q, H_p)}{\Pr(G_C|G_M, G_P, G_Q, H_d)\Pr(G_X|H_d)}$$

To make this example concrete, suppose the genotypes are

$$\begin{aligned}\mathrm{C}hild: &\quad G_C = A_1A_2 \\ \mathrm{S}ample: &\quad G_X = A_1A_3 \\ \mathrm{S}pouse: &\quad G_M = A_2A_4 \\ \mathrm{M}other: &\quad G_P = A_1A_5 \\ \mathrm{F}ather: &\quad G_Q = A_3A_6\end{aligned}$$

Then the probabilities need for the likelihood ratio are

$$\begin{aligned}\Pr(G_C|G_M, G_X, H_p) &= 1/4 \\ \Pr(G_X|G_P, G_Q, H_p) &= 1/4 \\ \Pr(G_C|G_M, G_P, G_Q, H_d) &= 1/8 \\ \Pr(G_X|H_d) &= 2p_1p_3\end{aligned}$$

so that

$$\mathrm{L}R = \frac{1}{4p_1p_3}$$

Exercise 6.4 Find the likelihood ratio for the situation when profiles are available from the mother, four siblings, the spouse, and a child of a missing person, as well as from a sample that may be from the missing person. Specifically,

$$\begin{aligned} \mathrm{M}other: &\quad G_P = A_3A_4 \\ \mathrm{S}ibs: &\quad \{G_S\} = \{A_2A_4, A_2A_4, A_2A_4, A_3A_4\} \\ \mathrm{S}pouse: &\quad G_M = A_5A_6 \\ \mathrm{C}hild: &\quad G_C = A_3A_5 \\ \mathrm{S}ample: &\quad G_X = A_3A_3 \end{aligned}$$

Use the fact that the genotypes of the four siblings and the mother imply that the father of the missing person must have had genotype A_2A_3 or A_2A_4.

Deceased Alleged Father

A similar scenario is when the alleged father in a paternity dispute cannot be typed, but typing is available from his relative(s). Suppose profiles are available from the mother and her child, allowing the paternal allele $A_P = A_i$ to be determined. The alleged father AF is deceased but his relative Z has been typed.

We need the probability that AF would transmit allele A_i when Z has a specified genotype. This is another instance where three-allele probabilities are needed. From the previous results for non-inbred individuals:

$$\begin{aligned} \Pr(A_P = A_i|G_z = A_iA_i) &= 2\theta_{AF,Z} + (1 - 2\theta_{AF,Z})p_i \\ \Pr(A_P = A_i|G_z = A_iA_j) &= \theta_{AF,Z} + (1 - 2\theta_{AF,Z})p_i, \;\; j \neq i \\ \Pr(A_P = A_i|G_z = A_jA_k) &= (1 - 2\theta_{AF,Z})p_i, \;\; j,k \neq i \end{aligned}$$

so the likelihood ratio is

$$\begin{aligned} \mathrm{L}R &= \frac{\Pr(A_M|G_M, H_p)\Pr(A_P|G_Z, H_p)\Pr(G_Z|H_p)}{\Pr(A_M|G_M, H_d)\Pr(A_P|G_Z, H_d)\Pr(G_Z|H_d)} \\ &= \frac{\Pr(A_P|G_Z, H_p)}{\Pr(A_P|H_d)} \\ &= \begin{cases} (1 - 2\theta_{AF,Z}) + \dfrac{2\theta_{AF,Z}}{p_i} & \mathrm{if}\; G_Z = A_iA_i \\ (1 - 2\theta_{AF,Z}) + \dfrac{\theta_{AF,Z}}{p_i} & \mathrm{if}\; G_Z = A_iA_j,\; j \neq i \\ (1 - 2\theta_{AF,Z}) & \mathrm{if}\; G_Z = A_jA_k,\; j,k \neq i \end{cases} \end{aligned}$$

If AF and Z are not related, $\theta_{AF,Z} = 0$, there is therefore no information in the genotype of Z concerning the paternity of AF, and the LR is 1 (Brenner 1997).

Table 6.9: LR values for an inheritance dispute.

G_X, G_Y	$\Pr(G_X, G_Y \mid H_p)$	$\Pr(G_X \mid H_d)\Pr(G_Y \mid H_d)$	LR
A_iA_i, A_iA_i	$\frac{1}{2}p_i^3(1+p_i)$	p_i^4	$\frac{1}{2}+\frac{1}{2p_i}$
A_iA_i, A_iA_j	$\frac{1}{2}p_i^2p_j(1+2p_i)$	$2p_i^3p_j$	$\frac{1}{2}+\frac{1}{4p_i}$
A_iA_i, A_jA_j	$\frac{1}{2}p_i^2p_j^2$	$p_i^2p_j^2$	$\frac{1}{2}$
A_iA_i, A_jA_k	$p_i^2p_jp_k$	$2p_i^2p_jp_k$	$\frac{1}{2}$
A_iA_j, A_iA_j	$\frac{1}{2}p_ip_j(4p_ip_j+$ $p_i+p_j)$	$4p_i^2p_j^2$	$\frac{1}{2}+\frac{1}{8p_i}+\frac{1}{8p_j}$
A_iA_j, A_iA_k	$\frac{1}{2}p_ip_jp_k(4p_i+1)$	$4p_i^2p_jp_k$	$\frac{1}{2}+\frac{1}{8p_i}$
A_iA_j, A_kA_l	$2p_ip_jp_kp_l$	$4p_ip_jp_kp_l$	$\frac{1}{2}$

Inheritance Dispute

Brenner (1997) discusses the following inheritance dispute: People X and Y have different mothers. The father Z of X has died, and Y claims also to be a child of Z. The two propositions for the genetic evidence of the genotypes of X and Y are

H_p: X and Y are half sibs.
H_d: X and Y are unrelated.

The likelihood ratio is

$$\begin{aligned} LR &= \frac{\Pr(G_X, G_Y \mid H_p)}{\Pr(G_X, G_Y \mid H_d)} \\ &= \frac{\Pr(G_X, G_Y \mid H_p)}{\Pr(G_X \mid H_d)\Pr(G_Y \mid H_d)} \end{aligned}$$

because the genotypes G_X and G_Y are independent when X and Y are unrelated. We now need to go back to Chapter 4 for the joint probabilities of genotypes of half sibs and list them in column 2 of Table 6.9. The joint probabilities in column 3 of Table 6.9 are just the products of the two separate probabilities. We have assumed no inbreeding and no population structure.

The likelihood ratios in Table 6.9 are less than one and against X and Y being half sibs unless they share rare alleles; for example for genotypes A_iA_j, A_iA_k we must have $p_i < 1/8$ for LR$>$ 1, and even then LR can decrease with more loci.

SUMMARY

Parentage testing, identification of remains, and inheritance disputes all exploit the genetic laws of transmission of alleles from parent to child. As with forensic applications, the DNA evidence is interpreted with likelihood ratios that compare the probabilities of the evidence under alternative propositions. Evaluation of the probabilities depends on using the laws of probability to make the probability of a genotype conditional on the parental genotype(s).

Chapter 7

Mixtures

INTRODUCTION

In Chapter 2 we considered the case in which the evidence at the crime scene consisted of two blood stains, and there was a single suspect whose genotype was the same as that of one of the stains. In this chapter we are going to extend that discussion considerably by talking about DNA profiles of samples that contain material from more than one contributor. The sensitivity of modern techniques is such that the incidence, complexity, and importance of such cases are increasing. It is not possible to present an exhaustive treatment of every eventuality, but by considering a range of different kinds of cases we hope to assist the reader in gaining a sufficient depth of understanding to tackle other situations as they arise. Rather than providing a recipe book, we adhere to the three principles for interpreting evidence listed in Chapter 2.

We will begin by considering the case in which independence of alleles within and between loci can reasonably be assumed, there are no population substructuring effects of practical magnitude, and all contributors to the mixed profile are from the same population (Evett et al. 1991). Later in the chapter we will relax these assumptions, but we will always assume that all contributors to the mixed profile are unrelated to each other, and that allelic dropout has no practical impact. This last assumption means that we will not be using the "2p" method that is widely used for the estimation of the frequency of single-banded RFLP-based VNTR systems (Weir et al. 1997). Moreover, we will carry out the analysis ignoring the intensities of electrophoretic bands or typing strip dots, or peak heights in histograms generated by automatic sequencers. We can refer to recent publications that do take into account peak heights (Evett et al. 1998). We adopt a simple

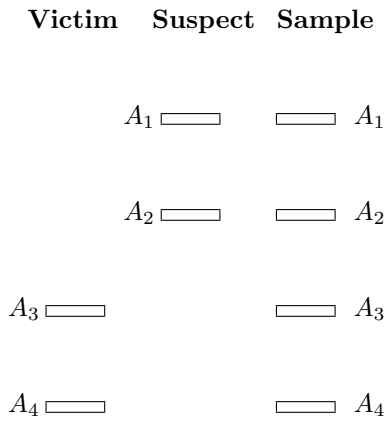


Figure 7.1: Four-allele mixture.

illustrative method for summarizing profiles: each allele will be represented by an open rectangle.

In our treatments of transfer evidence so far, we have always been able to consider just two hypotheses to explain the evidence: the prosecution and defense alternatives H_p and H_d respectively. This means that the odds form of Bayes' theorem can be used and the DNA evidence, summarized by the likelihood ratio, can be maintained distinct from the non-DNA evidence that is summarized by the prior odds. Later in this chapter we will meet situations in which there are more than two hypotheses. We will discuss the difficulties that arise and suggest methods for interpreting such cases.

VICTIM AND SUSPECT

The interpretation of a mixture depends very much on the circumstances surrounding the crime. For our first examples we envisage a case where there is very good reason for the victim's DNA to be present in the sample, as in a vaginal sample taken for alleged rape (ignoring the possibility of differential extraction of sperm and vaginal epithelial cell DNA). The crime sample therefore contains DNA from the victim and the offender, and is found to have both the victim's and the suspect's alleles. We also assume the victim alleges that no other man's DNA could be present in the sample.

Four-Allele Mixture

Suppose the DNA profiles for victim, suspect, and crime sample are as shown in Figure 7.1. The sample has four alleles, two of which match those of the victim and two of which match those of the suspect. The prosecution proposition is

> H_p: The crime sample contains DNA from the victim and the suspect.

We anticipate a defense proposition of the kind

> H_d: The crime sample contains DNA from the victim and an unknown person.

In Chapter 2 we used G_C to denote the genotype of a crime sample. That notation is not adequate for mixtures, and instead we now use E_C for the entire sample profile and G_V and G_S for the genotypes of victim and suspect, respectively. Omitting the non-DNA evidence I from the conditioning, purely for brevity, the likelihood ratio for the mixture is

$$\begin{aligned} \mathrm{L}R &= \frac{\Pr(E_C, G_V, G_S|H_p)}{\Pr(E_C, G_V, G_S|H_d)} \\ &= \frac{\Pr(E_C|G_V, G_S, H_p)}{\Pr(E_C|G_V, G_S, H_d)} \times \frac{\Pr(G_V, G_S|H_p)}{\Pr(G_V, G_S|H_d)} \end{aligned}$$

Because there is nothing in H_p or H_d that affects our uncertainty about G_V and G_S, the second ratio is one, and

$$\mathrm{L}R = \frac{\Pr(E_C|G_V, G_S, H_p)}{\Pr(E_C|G_V, G_S, H_d)}$$

If we now make the assumption, similar to Equation 2.4, that knowledge of G_S does not influence our uncertainty about the genotype of the offender, we then have

$$\mathrm{L}R = \frac{\Pr(E_C|G_V, G_S, H_p)}{\Pr(E_C|G_V, H_d)} \tag{7.1}$$

The numerator is one because the crime sample profile is exactly as expected if H_p is true. The denominator is the probability that an unknown person, unrelated to the victim, would contribute alleles A_1A_2 to the sample. As we are assuming here that this probability does not depend on the genotype of the suspect, it is $2p_1p_2$ under the Hardy-Weinberg assumption. Therefore

$$\mathrm{L}R = \frac{1}{2p_1p_2}$$

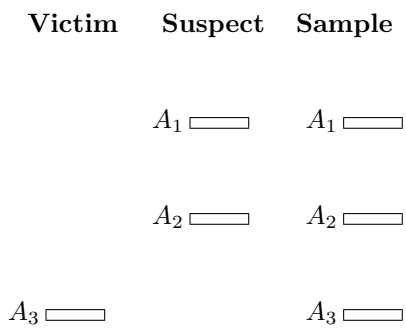


Figure 7.2: Three-allele mixture, victim homozygous.

This result establishes the common procedure of "subtracting" the victim's genotype from the crime sample profile.

Three-Allele Mixture

If the two contributors to a mixed stain have an allele in common, the profile of the stain will show only three alleles. We distinguish the cases of the victim being homozygous or heterozygous.

Victim homozygous. If the victim is homozygous for allele A_3 (Figure 7.2), the same line of argument as in the four-allele example leads to the same result:

$$\mathrm{L}R \;=\; \frac{1}{2p_1p_2}$$

Victim heterozygous. If the victim is heterozygous A_2A_3 and the suspect is homozygous A_1 (Figure 7.3), the numerator of the likelihood ratio remains equal to one, but the denominator requires more thought. It is necessary to consider three possibilities for the genotype G of the unknown person. The possibilities will be indexed by i:

i	G_i
1	A_1A_1
2	A_1A_2
3	A_1A_3

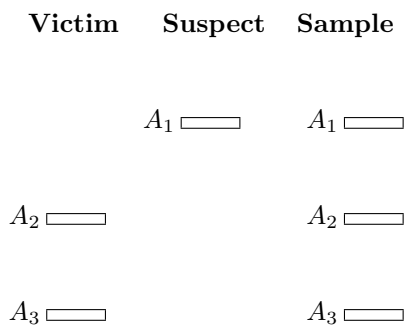


Figure 7.3: Three-allele mixture, victim heterozygous.

To evaluate the denominator of Equation 7.1, we apply the law of total probability (Chapter 1):

$$\begin{aligned}\Pr(E_C|G_V, H_d) &= \sum_i \Pr(E_C|G_V, G_i, H_d)\Pr(G_i|G_V, H_d) \\ &= \sum_i \Pr(E_C|G_V, G_i, H_d)\Pr(G_i|H_d)\end{aligned}$$

Because we are ignoring intensity differences, any one of the G_i plus the victim's genotype will lead to the crime sample profile, so

$$\Pr(E_C|G_V, G_i, H_d) = 1, \;\; i = 1, 2, 3$$

and then

$$\Pr(E_C|G_V, H_d) = \sum_i \Pr(G_i|H_d)$$

The probabilities of the G_i do not depend on H_d and are given by the products of allele frequencies. The likelihood ratio becomes

$$\mathrm{L}R = \frac{1}{p_1^2 + 2p_1p_2 + 2p_1p_3}$$

This likelihood ratio is less than the value $1/p_i^2$, which would have resulted if the suspect had the same genotype as a single-contributor stain of type A_1A_1. The presence of the victim's bands, in effect, has weakened the strength of the evidence against the suspect because they have increased the number of alternative hypotheses for the evidence if H_d is true.

Exercise 7.1 Find the likelihood ratio for the following cases in which the victim and offender are the only contributors to a stain, if it is known that the stain is from two contributors:

Victim	Suspect	Sample
A_1A_1	A_1A_1	A_1A_1
A_1A_2	A_1A_2	A_1A_2
A_1A_1	A_1A_2	A_1A_2
A_1A_2	A_1A_1	A_1A_2
A_1A_1	A_2A_2	A_1A_2

SUSPECT AND UNKNOWN PERSON

Some crime samples will contain DNA from more than one person, but only one known person is suspected of being a contributor. As before, we separate the cases of the sample showing three or four alleles.

Four-Allele Mixture

As in Chapter 2, when we considered the case of two stains at the crime scene, the hypotheses for a four-allele mixture profile that includes the genotype of a single suspect (Figure 7.4) are

> H_p: The crime sample contains DNA from the suspect and an unknown person.
> H_d: The crime sample contains DNA from two unknown people.

Following a similar line of reasoning as in the case where the victim was a contributor, we can show that

$$\mathrm{L}R \;=\; \frac{\Pr(E_C|G_S, H_p)}{\Pr(E_C|G_S, H_d)}$$

The genotypes G_i that might be components of the crime sample profile are

i	G_i
1	A_1A_2
2	A_1A_3
3	A_1A_4
4	A_2A_3
5	A_2A_4
6	A_3A_4

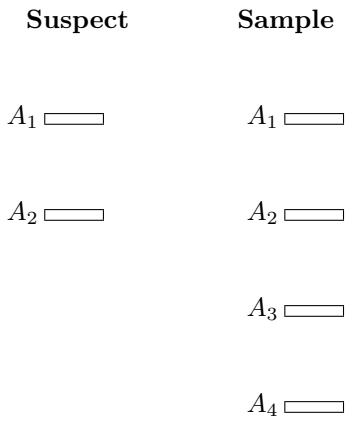


Figure 7.4: Four-allele mixture.

and we note that $G_1 = G_S$. The law of total probability gives

$$\Pr(E_C|G_S, H_p) = \sum_i \Pr(E_C|G_S, G_i, H_p) \Pr(G_i|G_S, H_p)$$

where G_i denotes the genotype of the unknown contributor to the mixture in addition to the suspect. This simplifies slightly to

$$\Pr(E_C|G_S, H_p) = \sum_i \Pr(E_C|G_1, G_i, H_p) \Pr(G_i|H_p)$$

where we have written G_1 for G_S. However, when we look at the possible values for G_i, we see that the evidence is not possible unless $i = 6$. So $\Pr(E_C|G_1, G_i, H_p) = 0$ when $i \neq 6$. The numerator for the likelihood ratio becomes

$$\begin{aligned} \Pr(E_C|G_S, H_p) &= \Pr(E_C|G_1, G_6, H_p) \Pr(G_6|H_p) \\ &= 1 \times 2p_3p_4 \end{aligned}$$

For the denominator of the likelihood ratio we first invoke our assumption that knowledge of G_S provides no information about the genotypes of possible contributors to the crime sample when H_d is true. Therefore

$$\Pr(E_C|G_S, H_d) = \Pr(E_C|H_d)$$

Because there are two unknown contributors under H_d, with genotypes G_i and G_j, the law of total probability gives

$$\Pr(E_C|H_d) = \sum_i \sum_j \Pr(E_C|G_i, G_j, H_d) \Pr(G_i, G_j, H_d)$$

The double summation indicates that we must consider every possible pair of genotypes from the list of permitted genotypes. Looking at that list, we see that there are only six combinations of i and j for which $\Pr(E_C|G_i, G_j, H_d)$ is not zero. They are all the possible pairs of heterozygotes that contain all four alleles $A_1A_2A_3A_4$ between them:

i	j	G_i	G_j	$\Pr(G_i, G_j\|H_d)$
1	6	A_1A_2	A_3A_4	$2p_1p_2 \times 2p_3p_4$
2	5	A_1A_3	A_2A_4	$2p_1p_3 \times 2p_2p_4$
3	4	A_1A_4	A_2A_3	$2p_1p_4 \times 2p_2p_3$
4	3	A_2A_3	A_1A_4	$2p_2p_3 \times 2p_1p_4$
5	2	A_2A_4	A_1A_3	$2p_2p_4 \times 2p_1p_3$
6	1	A_3A_4	A_1A_2	$2p_3p_4 \times 2p_1p_2$

We are still assuming that contributors i and j are unrelated, so that G_i and G_j are independent, and that any intensity differences are ignored. For each of the six combinations $\Pr(E_C|G_i, G_j, H_d) = 1$, and the likelihood ratio is

$$\begin{aligned} \mathrm{L}R &= \frac{2p_3p_4}{24p_1p_2p_3p_4} \\ &= \frac{1}{12p_1p_2} \end{aligned}$$

The likelihood ratio is reduced by a factor of six over what it would be if a suspect of type A_1A_2 were included in a single stain of type A_1A_2. If alleles A_1, A_2 were common in the population (e.g. $p_1 = p_2 = 0.3$ giving $12p_1p_2 > 1$) the likelihood ratio is actually less than one, meaning that the sample profile is more likely to be of type $A_1A_2A_3A_4$ if it came from two unknown people than if it came from the suspect and one unknown person. The evidence favors the defense, and this is one reason it is important to use the principles of evidence interpretation instead of simplistic rules of the "random man not excluded" type.

Three-Allele Mixture

Suspect heterozygous. When the suspect is heterozygous and the crime sample has those two alleles plus one other, as in Figure 7.5, the likelihood ratio is

$$\mathrm{L}R = \frac{\Pr(E_C|G_S, H_p)}{\Pr(E_C|G_S, H_d)}$$

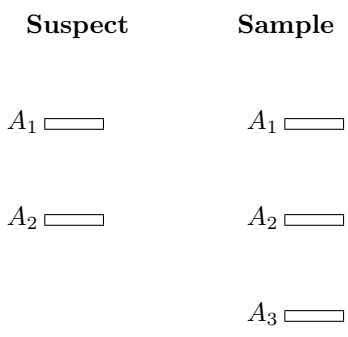


Figure 7.5: Three-allele mixture, suspect heterozygous.

and the genotypes that might be components of the crime sample profile are

i	G_i	$\Pr(G_i\|H_p)$
1	A_1A_2	$2p_1p_2$
2	A_1A_3	$2p_1p_3$
3	A_2A_3	$2p_2p_3$
4	A_1A_1	p_1^2
5	A_2A_2	p_2^2
6	A_3A_3	p_3^2

Note that $G_1 = G_S$. The numerator of the likelihood ratio is

$$\begin{aligned}\Pr(E_C|G_S, H_p) &= \sum_i \Pr(E_C|G_S, G_i, H_p)\Pr(G_i|G_S, H_p) \\ &= \sum_i \Pr(E_C|G_1, G_i, H_p)\Pr(G_i|H_p)\end{aligned}$$

Ignoring any intensity differences, only $i = 2, 3, 6$ allow the crime sample to have profile $A_1A_2A_3$, and these all give $\Pr(E_C|G_1, G_i, H_p) = 1$. Adding terms

$$\Pr(E_C|G_S, H_p) = 2p_1p_3 + 2p_2p_3 + p_3^2$$

For the denominator, there are 12 combinations of two genotypes that between them have the same profile as the crime sample. These are shown in Table 7.1. The denominator of the likelihood ratio simplifies to

$$\Pr(E_C|G_S, H_d) = 12p_1p_2p_3(p_1 + p_2 + p_3)$$

Table 7.1: Pairs of two genotypes with three alleles.

i	j	G_i	G_j	$\Pr(G_i, G_j \mid H_d)$
1	2	A_1A_2	A_1A_3	$2p_1p_2 \times 2p_1p_3$
1	3	A_1A_2	A_2A_3	$2p_1p_2 \times 2p_2p_3$
1	6	A_1A_2	A_3A_3	$2p_1p_2 \times p_3^2$
2	1	A_1A_3	A_1A_2	$2p_1p_3 \times 2p_1p_2$
2	3	A_1A_3	A_2A_3	$2p_1p_3 \times 2p_2p_3$
2	5	A_1A_3	A_2A_2	$2p_1p_3 \times p_2^2$
3	1	A_2A_3	A_1A_2	$2p_2p_3 \times 2p_1p_2$
3	2	A_2A_3	A_1A_3	$2p_2p_3 \times 2p_1p_3$
3	4	A_2A_3	A_1A_1	$2p_2p_3 \times p_1^2$
4	3	A_1A_1	A_2A_3	$p_1^2 \times 2p_2p_3$
5	2	A_2A_2	A_1A_3	$p_2^2 \times 2p_1p_3$
6	1	A_3A_3	A_1A_2	$p_3^2 \times 2p_1p_2$

and the ratio is

$$\mathrm{L}R \;=\; \frac{2p_1 + 2p_2 + p_3}{12p_1p_2(p_1 + p_2 + p_3)}$$

Exercise 7.2 For a crime sample of type A_1, A_2, A_3, known to contain DNA from two contributors, evaluate the likelihood ratio in the case where a suspect is of type A_2A_2.

TWO SUSPECTS

There are situations where the genotypes of two suspects are included in the profile of a mixture. We must repeat our frequent assertion that interpretation of a case such as this depends on the circumstances surrounding the crime. For this example, we assume that there is evidence that the crime was committed by two offenders. This could be the case in a double rape where the crime sample was the male fraction DNA extracted from a vaginal swab. Two suspects for the offense have been arrested for reasons that we have included in the non-DNA evidence I. We assume that the two suspects are to be tried together, in which case the prosecution proposition may be assumed to be of the kind

H_p: The crime sample contains DNA from the two suspects.

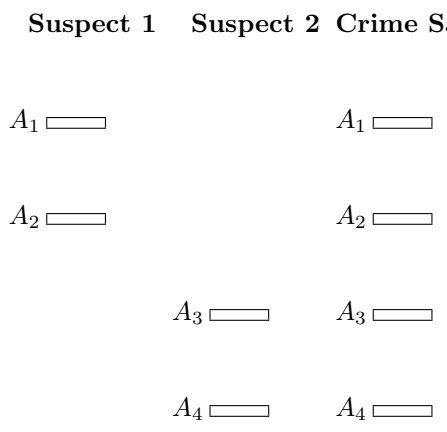


Figure 7.6: Banding pattern for four-allele mixture.

However, it now seems unrealistic to anticipate a simple defense proposition, particularly in view of the consideration that the two suspects will most probably be represented by different counsel. We can envisage three alternative defense hypotheses:

> H_{d1}: The crime sample contains DNA from suspect 1 and an unknown person.
> H_{d2}: The crime sample contains DNA from suspect 2 and an unknown person.
> H_{d3}: The crime sample contains DNA from two unknown people.

Four-Allele Mixture

In Figure 7.6 we show the situation in which the genotypes of two heterozygous suspects are included in the profile of a mixture. From the results of the previous sections, and continuing the same basic independence assumptions for the propositions just enumerated:

$$\begin{aligned} \Pr(E_C|H_p) &= 1 \\ \Pr(E_C|H_{d1}) &= 2p_3p_4 \\ \Pr(E_C|H_{d2}) &= 2p_1p_2 \\ \Pr(E_C|H_{d3}) &= 24p_1p_2p_3p_4 \end{aligned}$$

We have seen that when there are two alternative hypotheses to be addressed the scientific evidence can be summarized neatly by the likelihood

ratio, and this can be kept distinct from prior probabilities. With more than two hypotheses, however, this clear separation is not readily achievable. It is necessary to use the general form of Bayes' theorem (Box 1.4) to calculate posterior probabilities. For the prosecution proposition, using the law of total probability,

$$\Pr(H_p|E_C) = \frac{\Pr(H_p)}{\Pr(H_p) + 2p_3p_4\Pr(H_{d1}) + 2p_1p_2\Pr(H_{d2}) + 24p_1p_2p_3p_4\Pr(H_{d3})}$$

The forensic scientist is in no position to evaluate this because the prior probabilities are not within his or her domain and they will not be quantified at any time during typical court proceedings.

A possible solution to this is for the scientist to address a pair of alternatives at a time and calculate the ratio of probabilities of the evidence given each member of the pair. This is still a likelihood ratio even though the alternatives are not exhaustive. The results could be expressed in a table, though very careful explanation will be needed.

	Numerator		
Denominator	H_p	H_{d1}	H_{d2}
H_{d1}	$1/(2p_3p_4)$		
H_{d2}	$1/(2p_1p_2)$	$p_3p_4/(p_1p_2)$	
H_{d3}	$1/(24p_1p_2p_3p_4)$	$1/(12p_1p_2)$	$1/(12p_3p_4)$

It is not clear that any further simplification of the evidence is possible, unless the background circumstances change. Of course, the table would be simplified if it later transpired that one of the suspects pled guilty.

VICTIM AND/OR SUSPECT

Previously we considered the case in which the victim's DNA would be expected to be found under both hypotheses H_p and H_d. There will be situations, however, where this not the case. The crime sample in a rape case may be a blood-and-semen stain from some bedding. If this sample profile includes the genotypes of both victim and suspect, as in Figure 7.1, and if the victim identifies the bedding, the prosecution proposition would be the same as before:

> H_p: The crime sample contains DNA from the victim and suspect.

Anticipating the defense proposition is quite difficult in this case, but a start may be made with

> H_{d1}: The crime sample contains the DNA from the victim and an unknown person.
> H_{d2}: The crime sample contains the DNA from the suspect and an unknown person.
> H_{d3}: The crime sample contains the DNA from two unknown people.

Once more, we face the problem that the likelihood ratio formulation can be used only if we restrict attention to two propositions at a time.

There is no denying that this analysis would be difficult to explain and report. An alternative approach is to treat the evidence in two stages. The forensic scientist could say that he or she has considered the following two hypotheses:

> The stain contains the DNA of the victim and an unknown person.
> The stain contains the DNA of two unknown people.

Evaluating the likelihood ratio for these two alternatives proceeds as in the previous sections. The scientist would go on to report "The evidence is LR times more probable if the first of these hypotheses is true than if the second is true." The judgment of whether or not the victim was raped on the bedding from which the stain was recovered depends not only on the DNA evidence but also on other circumstances that the court will take into account. If the court determines that the bedding is indeed associated with the crime, then it may be meaningful to consider the following two hypotheses:

> The stain contains the DNA of the victim and the suspect.
> The stain contains the DNA of the victim and an unknown person.

The evaluation of the likelihood ratio for these two alternatives proceeds as before. The scientist concludes by saying "The evidence is LR times more probable if the first of these alternatives is true."

GENERAL APPROACH

The previous sections lay out a method for approaching the interpretation of specific mixed stains, but it is also helpful to have a general formula that makes the process somewhat automatic. This reduces the chance of failing to account for some of the possible genotypes for unknown contributors, and it allows the construction of computer programs.

The essence of the general approach is to identify the alleles in the crime sample and the alleles carried by known contributors to the sample. Any alleles in the sample not carried by known contributors must be carried by unknown contributors, and it is necessary to specify the number of unknown contributors. We also take into account the alleles carried by individuals who are known not to contribute to the sample. Note that the word "contributor" is defined by the proposition being considered–a person may be declared a contributor under one proposition and a non-contributor under another proposition. Evaluation of the probabilities needed for the likelihood ratio also requires knowledge of the number of times each allele occurs. We have set up the notation shown in Table 7.2 to help keep track of alleles and allele counts.

We write the set of all alleles in the evidence sample as $\mathcal{C}$. With n_C contributors, declared or unknown, there are $2n_C$ alleles in the set. However, not all the alleles will be different because some contributors may be homozygous and some may share alleles. We write the set of distinct evidence alleles as $\mathcal{C}_g$ and the number of distinct alleles as c. The whole set $\mathcal{C}$ contains c_i copies of allele A_i, $\sum_i c_i = 2n_C$. At present we are assuming allelic independence, so the probability of this set of alleles involves the product of probabilities for each allele in the set: $\prod_i p_i^{c_i}$. To complete the probability calculations we need to know the genotypic composition of the contributors to the sample. For a proposition H in which every contributor is specified, the genotypes are known and we need a factor of 2 for each of the h_C heterozygotes. The probability of the evidence is

$$\Pr(\mathcal{C}|H) \;=\; 2^{h_C} \prod_i p_i^{c_i}$$

However, at least one of the alternative propositions will involve unknown people, and then we may not know the genotypes of all contributors to the crime sample: we may not know the c_i's or h_C. Instead we have a set of n_T declared contributors, with allele counts t_i and h_T heterozygotes among them. There is also a set of n_V people declared not to be contributors, and these people have allele counts v_i and h_V heterozygotes among them. The x unknown people have u_i copies of allele A_i among them, $\sum_i u_i = 2x$

Table 7.2: Notation for mixture calculations.

Alleles in the profile of the evidence sample.	
$\mathcal{C}$	The set of alleles in the evidence profile.
$\mathcal{C}_g$	The set of distinct alleles in the evidence profile.
n_C	The known number of contributors to $\mathcal{C}$.
h_C	The unknown number of heterozygous contributors.
c	The known number of distinct alleles in $\mathcal{C}_g$.
c_i	The unknown number of copies of allele A_i in $\mathcal{C}$. $1 \le c_i \le 2n_C$, $\sum_{i=1}^{c} c_i = 2n_C$
Alleles from typed people that H declares to be contributors.	
$\mathcal{T}$	The set of alleles carried by the declared contributors to $\mathcal{C}$.
$\mathcal{T}_g$	The set of distinct alleles carried by the declared contributors.
n_T	The known number of declared contributors to $\mathcal{C}$.
h_T	The known number of heterozygous declared contributors.
t	The known number of distinct alleles carried by n_T declared contributors.
t_i	The known number of copies of allele A_i in $\mathcal{T}$. $0 \le t_i \le 2T$, $\sum_{i=1}^{s} t_i = 2T$.
Alleles from unknown people that H declares to be contributors.	
$\mathcal{U}$	The set of alleles carried by the unknown contributors to $\mathcal{C}$.
x	The specified number of unknown contributors to $\mathcal{C}$: $n_C = n_T + x$.
$c - t$	The known number of alleles that are required to be in $\mathcal{U}$.
r	The known number of alleles in $\mathcal{U}$ that can be any allele in $\mathcal{C}_g$, $r = 2x - (c - t)$.
r_i	The unknown number of copies of A_i among the r unconstrained alleles in $\mathcal{U}$. $0 \le r_i \le r$, $\sum_{i=1}^{c} r_i = r$.
u_i	The unknown number of copies of A_i in $\mathcal{U}$: $c_i = t_i + u_i$, $\sum_{i=1}^{c} u_i = 2x$. If A_i is in $\mathcal{C}$ but not in $\mathcal{T}$, $u_i = r_i + 1$. If A_i is in $\mathcal{C}$ and also in $\mathcal{T}$, $u_i = r_i$.
Alleles from typed people that H declares to be noncontributors.	
$\mathcal{V}$	The set of alleles carried by typed people declared not to be contributors to $\mathcal{C}$.
n_V	The known number of people declared not to be contributors to $\mathcal{C}$.
h_V	The known number of heterozygous declared noncontributors.
v_i	The known number of copies of A_i in $\mathcal{V}$. $\sum_i v_i = 2n_V$.

and $c_i = t_i + u_i$, but we may not know the genotypic composition of the x people. We can say that there are $(2x)!/\prod_i u_i!$ ways of arranging the alleles into genotypes, and this takes care of the factors of 2 for heterozygotes. For the counts u_i, the probability of the evidence is now

$$\Pr(\mathcal{C}, \mathcal{V}|H) \quad = \quad \frac{2^{h_T+h_V}(2x)!}{\prod_i u_i!} \prod_i p_i^{t_i+u_i}$$

It remains to assign values to the unknown counts u_i. The unknown people are constrained to carry at least one copy of the c distinct alleles in the crime sample that are not among the t distinct alleles carried by the declared contributors to the sample. This accounts for $c - t$ of the $2x$ alleles among the unknowns. Otherwise, they may carry any of the alleles in $\mathcal{C}_g$ but may not carry any allele not in $\mathcal{C}_g$. There are $r = 2x - (c - t)$ alleles in this unconstrained set, for which we write r_i for the number of A_i alleles. Note that r is known but the r_i are not known. For an allele in $\mathcal{C}_g$ but not in $\mathcal{T}_g$ we have that $u_i = r_i + 1$, and for the alleles in both $\mathcal{C}_g$ and $\mathcal{T}_g$ we have $u_i = r_i$. For any other allele, $u_i = 0$. It is a straightforward procedure to have a computer assign values to the r_is:

- Let r_1 take each value in the range $0, 1, \ldots, r$.
- Let r_2 take each value in the range $0, 1, \ldots, r - r_1$.
- ...
- Let r_{c-1} take each value in the range $0, 1, \ldots, r - r_1 - \ldots - r_{c-2}$.
- Then $r_c = r - r_1 - r_2 - \ldots - r_{c-1}$.

We can write the probability of the distinct alleles $\mathcal{C}_g = (A_1, A_2, \ldots, A_c)$ in the crime sample, for a proposition that has a set of alleles $\mathcal{T}$ carried by n_T declared contributors, a set $\mathcal{U}$ carried by x unknown contributors, and a set $\mathcal{V}$ carried by n_V known noncontributors, as

$$P_x(\mathcal{T}, \mathcal{U}, \mathcal{V}|\mathcal{C}_g) \quad = \quad \sum_{r_1=0}^{r} \sum_{r_2=0}^{r-r_1} \cdots \sum_{r_{c-1}=0}^{r-r_1-\ldots-r_{c-2}} \frac{2^{h_T+h_V}(2x)!}{\prod_i u_i!} \prod_{i=1}^{c} p_i^{t_i+u_i+v_i}$$

As an example, suppose the crime sample has alleles A_1, A_2, and A_3. Three people have been typed, and found to have the genotypes A_1A_1, A_2A_3, and A_3A_3. Such an example, for locus D1S80, was discussed by Weir et al. (1997). Two alternative propositions are

H_p: The crime sample is from the three typed people.
H_d: The crime sample is from three unknown people.

Under H_p, there are no unknown contributors and no noncontributors so, writing the empty set of alleles from unknown contributors or noncontributors as ϕ,

$$P_0(\mathcal{T} = A_1A_1A_2A_3A_3A_3, \mathcal{U} = \phi, \mathcal{V} = \phi | A_1A_2A_3) \quad = \quad 12p_1^2p_2p_3^3$$

and this is $\Pr(E|H_p)$, the probability of the evidence under H_p.

Under H_d, there are no declared contributors and there are three unknown contributors. The unknown contributors must carry alleles $A_1A_2A_3$, leaving $r = 3$ unconstrained alleles, and $u_i = r_i + 1, i = 1, 2, 3$. There are also three declared noncontributors, who have alleles $A_1A_2A_3$ with counts $2, 1, 3$ and one heterozygote. The probability, for $\mathcal{T} = \phi, \mathcal{U} = A_1A_2A_3, \mathcal{V} = A_1A_2A_3$, is

$$P_3(\phi, \mathcal{U}, \mathcal{V} | A_1A_2A_3) \quad = \quad \sum_{r_1=0}^{3} \sum_{r_2=0}^{3-r_1} \frac{2^1 p_1^{r_1+1} p_2^{r_2+1} p_3^{4-r_1-r_2}}{(r_1+1)!(r_2+1)!(4-r_1-r_2)!}$$

There are 10 terms in the summation, corresponding to r_1, r_2, and r_3 values of (0,0,3), (0,1,2), (0,2,1), (0,3,0), (1,0,2), (1,1,1), (1,2,0), (2,0,1), (2,1,0), and (3,0,0). The 10 add to the probability of the evidence under H_d:

$$\begin{aligned} \Pr(E|H_d) \quad = \quad & 360p_1^3p_2^2p_3^4[p_3^3 + 2p_2p_3^2 + 2p_2^2p_3 + p_2^3 + 2p_1p_3^2 \\ & + 3p_1p_2p_3 + 2p_1p_2^2 + 2p_1^2p_3 + 2p_1^2p_2 + p_1^3] \end{aligned}$$

and the likelihood ratio is the ratio $\Pr(E|H_p)/\Pr(H_d)$.

Effects of Population Structure

The general approach in the previous section assumed that all the alleles were independent. That can be modified very simply (Curran et al. 1999) to allow for the kinds of dependence imposed by joint membership in the same subpopulation. The Dirichlet theory described in Chapter 4 is appropriate, and Equation 4.23 is needed. Recall that this theory does assume independence of alleles within the subpopulation, but dependence in the whole population. The theory supposes that allele proportions are available only for the whole population. The quantity θ serves to quantify the variation of allele proportions among the subpopulations, and allows the population-wide proportions to be used for a subpopulation. The term $\prod_i p_i^{t_i+u_i+v_i}$ in the general expression of the last section is replaced by

$$\frac{\Gamma(\gamma_.)}{\Gamma(\gamma_. + 2n_T + 2x + 2n_V)} \prod_{i=1}^{c} \frac{\Gamma(\gamma_i + t_i + u_i + v_i)}{\Gamma(\gamma_i)}$$

with $\gamma_i = (1-\theta)p_i/\theta$ and $\gamma_\cdot = \sum_{i=1}^{c} \gamma_i$ as before.

It is possible to refine this to allow some of the people to be in one subpopulation, and some in another (Curran et al. 1999). This would allow a victim and her attacker to be in different subpopulations, or even different racial groups, but the suspect and attacker to be in the same subpopulation. It is necessary to apply separate Dirichlet moment formulations to each different subpopulation and also to account for the orderings of alleles among individuals separately for each subpopulation. This refinement may not be necessary because assigning all alleles to the same subpopulation maximizes the effects of population substructure, and separate analyses can be made with allele proportions from different racial groups.

NUMBER OF CONTRIBUTORS

The analyses so far have all assumed that the number of contributors to a mixed sample is known. In some cases this will be a reasonable assumption, but in other cases there may be little information about the number of unknown contributors. A complete analysis would allow for different numbers of unknown contributors, each number with its own prior probability. However, these priors are likely to be outside the province of the forensic scientist. An alternative is to provide separate analyses for each of a range of numbers of unknown contributors. By and large, for a locus that has all possible alleles present in the mixture, the probability of the mixture profile increases with the number of contributors. It becomes more probable that a large number of contributors will have all the alleles at a locus between them. The opposite is true for a locus where only some of the possible alleles are present in the mixture profile. Then it becomes less probable that a large number of contributors will have only that set of alleles between them.

SUMMARY

The interpretation of mixed stains is possible only in the context of likelihood ratios. Unlike single-contributor stains, the sample profile may not be certain under either of two alternative propositions, so the likelihood ratio is the ratio of two probabilities that are less than one. Presenting the probability under only one proposition can be quite misleading.

Throughout this chapter we have ignored any information, such as band intensity, that may indicate the relative amounts of DNA from different contributors in a mixed sample. Taking account of intensity differences can increase discrimination (Evett et al. 1998), but at present the statistical

methodology is at a comparatively early stage. There are also semi-intuitive methods that discount selected genotypic combinations, but these methods must be used with extreme care as they can give misleading results.

Chapter 8

Calculating Match Probabilities

INTRODUCTION

In previous chapters we covered the basic elements of evidence interpretation: probability, statistics, population genetics, and statistical genetics. We also treated the special needs of parentage testing and interpreting DNA mixtures. In this chapter we summarize the procedures for calculating match probabilities, and in the final chapter we cover the issues involved in presenting evidence. We phrase the discussion in this chapter in terms of the Gotham City example we first met in Chapter 3.

Using the terminology we developed in Chapter 2, we denote the genotype of the crime stain as G_C and that of the suspect as G_S. The two genotypes have matched, in other words $G_C = G_S = G$. As in Chapter 2, we consider two propositions:

H_p: The suspect left the crime stain.
H_d: Some other person left the crime stain.

In Chapter 3 we considered how a database of profiles from a Gotham City convenience sample could be used to assist in weighing these alternatives against each other. In particular, we showed how the database could be used to estimate allele proportions for each of the alleles in the matching profile. We now consider the discussion of how the strength of the evidence should be evaluated.

PROFILE PROBABILITY

The evaluation of the LR depends greatly on what we mean by the "some other person" who is the true offender, given that H_d is true. If, given H_d, we can assume the suspect and offender are unrelated, then we can drop G_S from the conditioning in the denominator. The LR is then simply the reciprocal of $\Pr(G_C|H_d, I)$, as in Equation 2.5, and this can be estimated as the genotype proportion in the whole population.

If we believe that it is reasonable to regard our convenience sample as a random sample (Chapter 3), the sample proportion is unbiased for the population proportion. This is just a property of the multinomial distribution, also discussed in Chapter 3. The problem is, as the number of loci in the profile increases, it becomes very unlikely that any particular profile will be found in our sample. It is necessary to construct an estimate, and one approach would be to form the product of allele proportions, as discussed in Chapter 5. This would be a valid procedure only if these proportions were independent.

We may be willing to assume independence if we could be sure of the homogeneity of the population from which the convenience sample was drawn. However, once we acknowledge the possibility of population substructuring, we know from Chapter 4 that independence cannot hold. The classical statistical approach is to proceed with independence testing in spite of the recognition that the null hypothesis of independence is not true–the thought being that the degree of dependence may be too low to cause the product rule to be misleading. Even then, we showed in Chapter 5 that tests are expected not to find departures from independence until those departures are quite large.

Problems with Independence Testing

There are a number of problems with independence testing, which we will now review. Recall that the basic steps of hypothesis testing, such as for within-locus or between-locus independence, can be summarized as follows:

- Set up the null hypothesis that independence is the true state.
- Calculate a test statistic from the data.
- Compare the value of the test statistic with the probability distribution (e.g., chi-square) that holds when the null hypothesis is true. Use that distribution to calculate the probability of observing a value equal to or greater than the test statistic. This is the P-value.

- If the P-value is small, reject the null hypothesis.

We now consider some issues arising from this procedure.

The null hypothesis. Recall that in Chapter 4 we explained how the theory of population genetics was based on simple models that invoked the idea of an ideal population. The assumptions that we adopted for the derivation of the Hardy-Weinberg law included the notion of a random-mating population of infinite size, although this is never realized in real-world populations. Not only are human populations of finite size, but they also do not mate at random because of many factors such as geography, religion, education, and socioeconomics. The conditions for complete independence do not exist for real human populations. It follows that the null hypothesis cannot be true in any practical situation.

The alternative hypothesis. The null hypothesis is absolutely precise: independence exists. In any scientific endeavor there must always be at least one alternative hypothesis to explain a phenomenon or state of nature. The alternative hypothesis in this case must be that independence does not exist. Note that, in contrast to the null hypothesis, the alternative hypothesis is infinitely vague. In fact, we already know that the conditions for the null hypothesis cannot exist, so even without collecting data, we know that the null hypothesis must be false and the alternative hypothesis must be true.

Classical statistics very often takes the "straw man" approach of setting up a null hypothesis suspected of being false and then examining data to see if indeed the hypothesis can be rejected. In the present context of independence testing, the desired outcome is a failure to reject. Part of the attraction for a Bayesian approach to statistical inference is that it avoids the hypothesis-testing approach and the difficulty of testing a null hypothesis known to be false against an alternative known to be true.

The P-value. The P-value is the probability, given the truth of the null hypothesis, of values of the test statistic equal to or more extreme than that which has been observed. The standard view is that small P-values cast doubt on the truth of the null hypothesis. What do we mean by "small"? An element of arbitrariness inevitably creeps in here: conventional wisdom is that a P-value of 0.05 is "significant" and one of 0.01 is "very significant." So if, for example, we arrive at a P-value of 0.005 we might say that we reject the null hypothesis at the 1% level, and proceed as though the null hypothesis were not true–which for the hypothesis of Hardy-Weinberg equilibrium is as it should be!

"Proving" independence. On the other hand, what if a moderate P-value is calculated–0.5, for example? It means that the null hypothesis is not rejected, but one must not fall for the misconception that the null hypothesis has therefore been proved to be true. That is most certainly not the case. All we have demonstrated is that the data do not provide sufficient evidence to *disprove* the null hypothesis. This may well happen when the data set is small. We can never *prove* independence; all we can ever do is to say that the dependence effects, whatever they are, were not detected by our test. This may be a consequence of the effects indeed being very small, or it may be a consequence of lack of data, or of the design of the test statistic.

"Proving" dependence. So, if we have a small P-value we reject the null hypothesis, which seems to suggest that the null hypothesis is false. We appear to have proved that there are dependence effects. The danger here is there has been a school of thought that holds the view that this means the product rule should not be used. But this is not necessarily the case: we must recognize that there is a clear difference between STATISTICAL significance and PRACTICAL significance. It might be that the quantity of data and the design of the test statistic are such that effects are manifested that, although real, are still far too small to have any noticeable impact on the figures put to a court for interpreting a DNA match. The important point here is that the P-value is not a measure of how well the estimation procedure will work in casework.

Multiple testing. As we employ more and more loci to aid our discrimination in DNA cases, so the potential numbers of independence tests grow. For example, with a six-locus STR multiplex there are six within-locus tests for each database , and the potential for false rejections will increase above the nominal value of 5% that we believe is the case, for example, when we reject on the basis of a chi-square test statistic exceeding 3.84. The Bonferroni correction (Weir 1996) for multiple tests makes a distinction between "comparison-wise" and "experiment-wise" significance levels. Our discussion on P-values so far has been comparison-wise. Each test, considered singly, has probability P of leading to false rejection. If, for example, we were carrying out six independent tests then the experiment-wise significance level is the proportion of times that one or more of all six tests would lead to a false rejection. If it is desired to keep the experiment-wise rate at level P, then the rate for each individual test needs to be decreased to (approximately) P/N when N tests are performed. In the present example, each of the six tests would be conducted at the $P = 0.008$ level, to give

an experiment-wise error rate of 5%. A chi-square statistic would need to be larger than 6.96 instead of 3.84 to indicate rejection. This approach is not really satisfactory, as the six tests are not replicates of the same test. We are really interested in departures from Hardy-Weinberg at each locus individually–it is not that all six single-locus tests are addressing exactly the same issue. The number of tests increases if we also examine sets of three, four, five, and six loci and the potential for false rejections would increase even if perfect independence existed–which it doesn't, as we have seen. Knowing that, what should we do when we get a significant rejection for a particular combination?

Prior knowledge. A serious weakness of the hypothesis-testing approach is that it does not enable account to be taken of prior knowledge. There may be very good prior reasons that dependence effects in a given population are minor. These could include previous studies at other loci; knowledge that the new loci of interest are evolutionarily neutral; demographic information showing that subpopulations have experienced generations of intermixing; or sociological studies showing that inbreeding levels are small.

Post-hoc rationalization. A natural consequence of the last two points is the practice, of which there are many examples to be found in the literature (e.g., Evett et al. 1996a) where one or more "significant" results in a study are discounted by the authors as being of no importance. The literature has demonstrated various approaches to doing this: carrying out additional tests based on different test statistics; localizing the genotypic combinations that contribute most to the test failure and demonstrating that the departure has small practical effect; citing other studies in which the effect was not observed; and invoking the fact that multiple testing has been carried out, implementing the Bonferroni inequality to weaken the power of all of them.

The issue of testing for independence across loci is even more complicated, and in Box 5.8 we referred to the great difficulty in performing meaningful tests for allelic independence at several loci.

Why should we employ hypothesis testing? A good proportion of this book has been devoted to hypothesis testing, and the reader can be forgiven for being puzzled by an apparent *volte-face* by the authors. There are indeed some good reasons for using a significance test, and we point out that tests have been used in many of the scientific advances of the 20th century. For example, new drugs are adopted after rejection of null

hypotheses that they have no beneficial effect or that they are no better than existing medications.

With much of existing practice steeped in this approach to scientific inference, it may be difficult to introduce a new system of genetic markers into casework unless it can be shown that a minimum standard battery of tests has been applied. Next, some of the tests are very simple to carry out, can be very useful for a first pass over the data, and can provide an early indication that some hitherto unexpected effect is operating–such as, for example, a technical problem leading to allelic drop-out and consequent excess homozygosity. Finally, we must admit that alternative methods can be mathematically far more complex than most hypothesis tests.

MATCH PROBABILITIES

The need for assuming independence is removed when we proceed as though there is always some degree of association between the genotypes of the suspect and the offender, and calculate the match probability $\Pr(G_C|G_S, H_d, I)$ and thus the LR at Equation 2.3. Our expressions for the match probability in Equations 4.20 allow for dependences due to population structure.

In Chapter 4 we developed a theory for these probabilities when, under H_d, the offender and suspect are related by virtue of being in the same family or by virtue of shared evolutionary history. For our Gotham City example, we can ignore the possibility of family relationships. We do want to consider evolutionary dependence, however, and the most conservative statement we can make about the unknown offender is that he belongs to the same subpopulation as the suspect. Because the convenience sample is from the whole population, our treatment of match probabilities took explicit account of population structure by means of the parameter θ (Equations 4.20):

$$\begin{aligned} \Pr(G_S = A_iA_i|G_C = A_iA_i) &= \frac{[2\theta + (1-\theta)p_i][3\theta + (1-\theta)p_i]}{(1+\theta)(1+2\theta)} \\ \Pr(G_S = A_iA_j|G_C = A_iA_j) &= \frac{2[\theta + (1-\theta)p_i][\theta + (1-\theta)p_j]}{(1+\theta)(1+2\theta)} \end{aligned} \tag{8.1}$$

Implicit in the development of these equations is the recognition that population structure produces allelic dependence in the whole population. In other words, use of these equations avoids the need to assume allelic independence in the whole population–indeed dependence is assumed. Use of these equations also avoids the need to specify the subpopulation since

the result is expected to hold for any subpopulation. The next issue is that of choosing an appropriate values for θ.

Choosing θ

In our Gotham City example we have just the one convenience sample from the whole population, so how can we decide on a value of θ to use? It is a mistake to be prescriptive, and we would urge scientists to consider each case on its own merits. Guidance can be sought from the literature; demographic studies are useful as well as analyses carried out on genotyping systems other than those in forensic use. We also bear in mind that the forensic scientist will, in general, desire to err on the conservative side.

The parameter θ refers to the relationship of pairs of alleles within a subpopulation relative to that between alleles in different subpopulations. It also serves as a measure of differences among subpopulations. The variance of allele proportions among subpopulations is proportional to θ. The ideal situation would be to have data from different subpopulations in order to estimate the θ values appropriate for each one, as described in Chapter 5. This is not practical, not least because of the difficulty in allocating people to subpopulations. There are two possible solutions. One is to refer to previous studies of human population structure, such as the monumental compilation of Cavalli-Sforza et al. (1994). Although these authors used different loci, they did study very many populations and we consider their results to be relevant. They reported θ estimates that were generally less than 0.05, which is in agreement with our understanding of human evolution and the graph we showed in Figure 4.7 for $N = 100,000$. The other solution is to adopt an arbitrary value of θ that could be considered conservative, as did the 1996 NRC report (National Research Council 1996). The report contains a useful discussion of this issue, and for STR systems, suggests values in the range 0.01 to 0.03 until practitioners acquire the appropriate data to carry out studies of structuring within their own environment. We support this recommendation.

It is worth stressing that we do not attempt to define a point at which the relationship among people gives a θ of zero. Our understanding of human evolution (e.g., Cavalli-Sforza 1998) is that *all humans are related*. This is a simple consequence of the fact that no two people currently alive can have had distinct sets of ancestors for all of the past 200,000 years. On the other hand, of course, it is for small subpopulations that people have had the longest shared history and so have the largest θ values. It seems appropriate to adopt a conservatively high value of θ with the understanding that it

would accommodate people of the same race in the same subpopulation as well as people in quite distinct races.

We discussed the classical methods for estimating θ in Chapter 5. Alternative Bayesian methods have been explored by Balding and Nichols (1997) and by Foreman et al. (1997).

Balding and Nichols modeled θ for the jth locus in the ith subpopulation by

$$\theta_{ij} = \frac{1}{1+\alpha_i+\beta_j}$$

and assigned lognormal prior distributions to α and β. Allele proportions were assumed to have a Dirichlet distribution. The authors used data from a geographic-ally-defined subpopulation and from a heterogeneous large population. Their estimate therefore referred to the relationship of alleles within each of the subpopulation and the large population when compared to alleles between the subpopulation and the large population.

Foreman et al. also allowed θ to vary over subpopulations and loci. They overcame the lack of subpopulation data by partitioning the population sample into arbitrary subsamples. Their likelihood assigned most weight to those subdivisions that were most plausible because they consisted of the individuals with the most similar DNA profiles. When the number of partitions is known, the authors were able to show that their methods give good estimates of θ. Their analysis was also Bayesian, with a Beta prior distribution for θ and a Dirichlet prior for the allele proportions p_i.

Both sets of authors found that posterior distributions for θ were skewed, with a long tail to the right. For the data they examined, however, the distribution pointed to values that rarely exceeded a few percent.

Practical Impact of the Procedure

We have seen how the procedure for calculating match probabilities enables us to take account of dependence effects within loci. Next we face the problem of combining our single locus match probabilities across loci. There is no comparable body of theory for taking account of between-locus effects, but we settle for simple multiplication over loci with the thought that increasing θ will take account of further dependence effects. Of course, we do not know how much to increase θ, and an element of arbitrariness is unavoidable, but our judgment can be informed by carrying out simple experiments on the data that enable us to assess the practical effects of using different values of θ.

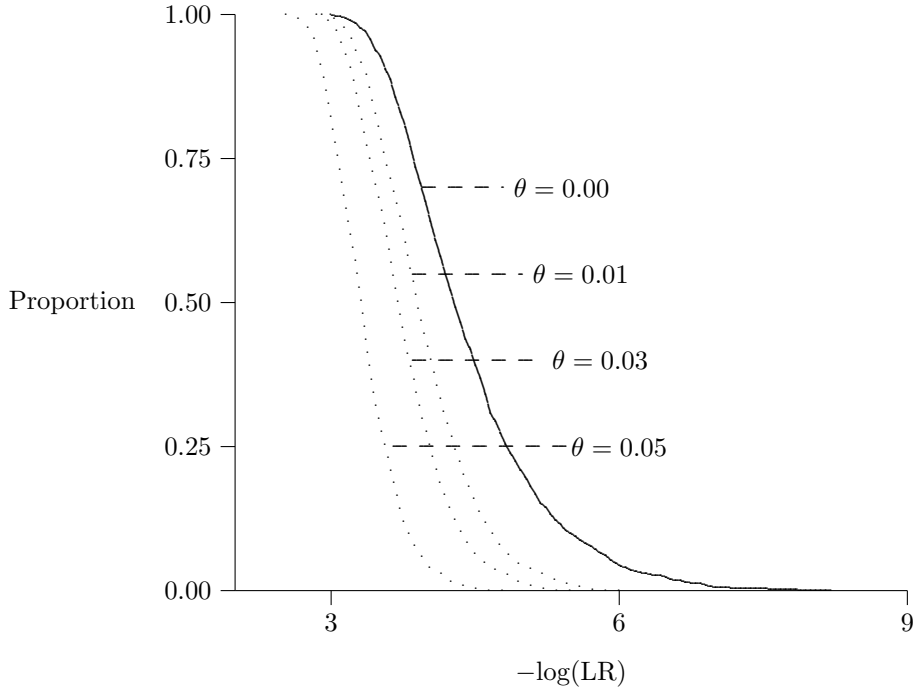


Figure 8.1: Within-person Tippett plot for UK Caucasian data.

First, we can simulate "within-person" comparisons (i.e., cases in which the offender and suspect are the same person) by calculating the LR for each person represented in a database. We can start by investigating the consequences of ignoring population structure altogether, calculating Hardy-Weinberg proportions $\hat{P}_{ii} = \hat{p}_i^2$ and $\hat{P}_{ij} = 2\hat{p}_i\hat{p}_j$ at each locus, and multiplying across loci. The LRs are then taken as the inverses of these probabilities, as in Equation 2.4. For a UK Caucasian quadruplex data set (Evett et al. 1996b), there are 1,401 LR values, and these lead to the solid line "Tippett plot" in Figure 8.1 (Evett and Buckleton 1996). The vertical scale shows the proportion of cases in which the LR exceeds the value on the horizontal scale (which is logarithmic). For example, in practically all cases in which the suspect is the offender, LR values in excess of 1,000 are expected for this set of four loci. In about 25% of such cases, LR values in excess of 100,000 are expected.

This Tippett plot follows from using assumptions of independence within and among loci. However, we recognize that independence does not hold, and we avoid the assumption (at least within loci for structured populations) by using Equations 8.1. The consequences of adopting these equations, with observed allele proportions substituted for the p_i values, and with θ set to

0.01, 0.03, and 0.05, are shown as dotted lines in Figure 8.1. Overall, use of Equations 8.1 reduces the LR values, and the effect is most pronounced for profiles that have alleles with small proportions, i.e., profiles with large LRs. The effects also increase as θ increases. We consider that $\theta = 0.03$ is a conservative upper bound on the values appropriate for human populations, as has been suggested previously (National Research Council 1996).

The analysis can next be taken a stage further by carrying out all the possible comparisons between pairs of profiles in the database. For a sample of size n, this gives $n(n-1)/2$ "between-person" comparisons. For the same sample of 1,401 Caucasians being discussed here, this means 980,700 comparisons. In most of them the genotypes will be different and the LR will be zero. Whenever two genotypes match, however, the LR can be calculated by multiplying together allele proportions within and between loci. The LR distribution for this experiment applied to the UK data is shown in Figure 8.2. Only 118 matching four-locus profiles were found. The solid line is the same kind of plot as for the previous experiment, but the vertical scale is quite different–it is in terms of matches per 100,000. So, based on this analysis, we estimate that a match between two unrelated people, and subsequent LR in excess of 1,000, will occur in approximately 12 cases per 100,000. The number of cases in which the LR will exceed 10,000 is about 1 in 50,000. The experiment gives us a measure of the discriminating power of the genotyping system, and it also allows a demonstration of the effects of applying Equations 8.1. Again, in Figure 8.2, the dotted lines show the effects of setting θ to 0.01, 0.03, and 0.05.

Multiple Loci

There remains the need for further study of the problem of assigning numerical values to Equations 8.1. One issue that is presently unresolved is that of how to accommodate multiple loci in a rigorous manner. We have suggested the multiplication over loci of LRs found from Equations 8.1, under the assumption that any dependencies will be small. We acknowledge, however, that strict independence is not true and that testing for allelic dependencies at multiple loci is difficult at best.

As a practical matter, we do not consider that multiplication over loci is in any way misleading, and we do believe that the LR for a matching profile is increased by having more loci in the profile. It is a matter of common experience that the proportion of between-person matches in large databases decreases steadily as the number of loci increases, and we note that most loci used for identification are unlinked.

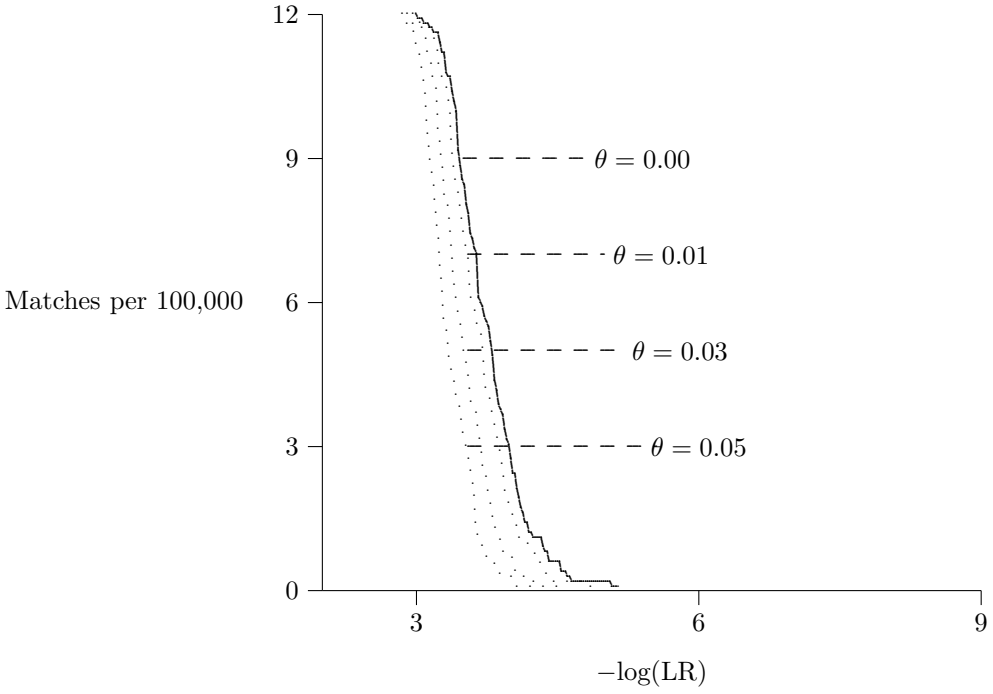


Figure 8.2: Between-person Tippett plot for UK Caucasian data.

SUMMARY

When profiles from a crime scene and a suspect are found to match, the likelihood ratio depends on the match probability, or the probability of the crime scene profile conditional on the suspect's profile. The two profiles are dependent, even for unrelated people, because of evolutionary history, and the relationship is described by Equations 8.1. Using these equations eliminates the need to test for allelic dependencies. Assigning numerical values is not straightforward, but a satisfactory approach is to use allele proportions from convenience samples and to use estimated or assigned values of θ.

Chapter 9

Presenting Evidence

INTRODUCTION

One view of forensic scientists is that they should carry out examinations and tests according to a documented protocol, and confine their reports and evidence to a statement of observations. We do not subscribe to this view, and instead believe that the scientists should have the understanding necessary to give rational, constructive, and balanced advice to a court on evidence interpretation. The role of the forensic scientist should be that of providing expert opinion, whether the evidence refers to DNA profiles or more traditional items such as fingerprints, glass fragments, or hair and fibers. Although we have explained methods in this book for attaching statistics to DNA profiles, we have tried to avoid giving the impression that the interpretation of DNA evidence is purely objective. We certainly embrace the notion that forensic science evidence must be objective in the sense of being impartial and not influenced by the prejudices that might influence the nonscientific aspects of evidence in a particular case. But we do not accept that DNA statistics are objective in the sense of being independent of human judgment. In spite of the often elegant mathematical arguments we have presented, we stress that the final statistical values depend wholly on the initial assumptions. The validity of these assumptions in any given case are a matter for expert opinion, so that we claim "objective science" can exist only within the framework of subjective judgment.

In this chapter we will explain our views about preparing and presenting an assessment of the evidence in a case in which there is a match between a crime profile and a suspect. We will then discuss the presentation of evidence in court with reference to issues arising from certain court judgments. We also discuss the subject of individualization.

As a focus for our discussion, we return to the example we introduced at the beginning of Chapter 2 and extended in Chapter 3. We elaborate further by describing the circumstances of the crime and we will discuss their relevance to the interpretation.

Ms. V. was walking home along a deserted street in a residential area of Gotham City in the early hours of the morning when she was attacked by a man and raped at knife point. The attacker made off, after making various threats about what would happen if she screamed before he had got well away. A passer-by later came to her assistance and the police were called. She was subsequently given a medical examination and vaginal swabs were taken. She described her attacker as Caucasian, around 1.8 m tall, with dark hair, medium build, in the age range 25 to 30 years. She said that he had spoken to her with a local accent. No suspect was immediately apparent to the investigator, but the vaginal swabs were submitted to the forensic laboratory for examination. Copious sperm heads were observed and a full DNA profile was obtained.

A few days later a police officer arrested a man, Mr. S., as a result of a lead from an informant. He provided a sample that, when profiled, was found to be the same genotype as that of the semen recovered from the vaginal swabs. There was no evidence that Mr. S. had any close male blood relatives in the city.

CALCULATION OF LIKELIHOOD RATIO

The circumstances of this case suggest that its interpretation follows the lines of the "single crime scene stain" case that we considered at the start of Chapter 2. The likelihood ratio is then the inverse of the match probability as in Equation 2.3. The circumstances also suggest that Mr. S. has no close relatives in the city, so the match probability for each locus is appropriately calculated using Equations 4.20. Note that, because the victim described her assailant as Caucasian, it is appropriate to use estimates of allele proportions from a Caucasian database. The expert also faces a decision about the choice of θ, and about how to combine the probabilities across loci. The second NRC committee (National Research Council 1996, hereafter referred to simply as the 1996 NRC report) considered these issues, and concluded that it was reasonable to multiply across loci. They also gave general guidance for values of θ, suggesting the range 0.01 to 0.03. We do not intend to give any further guidance ourselves because we believe that this is a matter for the judgment of the scientist, to be taken within the circumstances of the particular case. We hope that he will have the results of studies to hand

based on appropriate data that will guide his judgment: in their absence, he will no doubt err on the side of caution and choose a fairly large value, such as 0.03.

We repeat our emphasis of the importance of the judgment of the scientist here. The central feature of the evaluation of the weight of the evidence is a *probability*, and we have already emphasized the personal nature of probability. We hope that we have done enough elsewhere in the book to show that there is no "right" number for this probability: the number that is given represents the scientist's reasoned and, in his opinion, balanced assessment of the weight of the evidence. Let us assume for the Gotham City example that the scientist calculates a match probability of one in a million, and therefore a likelihood ratio of one million.

RESULT OF A DATABASE SEARCH

In our example, we have explained that Mr. S. became a suspect for the rape of Ms. V. because of a lead from an informant. There is another kind of situation in which a suspect comes to notice: his profile is stored on a database of previous offenders. Should the fact that a suspect came to notice from a database search affect the evaluation of the weight of the scientific evidence? Although that is not the case in our example, this is a convenient point to discuss the question.

Let us imagine that the forensic science laboratory in Gotham City maintained a database of the profiles of previous offenders and that Mr. S. had come to notice because the profile from the vaginal swab was searched against the database and his profile was found to match. Let us assume that the database contained the profiles of N men and that Mr. S.'s was the only one to match. How would this affect the evaluation of the evidence? The NRC committee considered this issue and made this recommendation (Recommendation 5.1):

> When the suspect is found by a search of DNA databases, the random-match probability should be multiplied by N, the number of persons in the database.

So, if the Gotham City database contained 10,000 men, the recommendation would mean that the scientist would give a likelihood ratio of 100, rather than of one million. This would be a drastic dilution of the strength of the evidence and we need to look at the underlying logic rather more closely.

Recall, from Chapter 2, that in a case in which the evidence consisted of a stain left at the crime scene, the likelihood ratio took the form

$$\mathrm{L}R \;=\; \frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)} \tag{9.1}$$

where $E = (G_S, G_C)$ and $G_C = G_S$. Balding and Donnelly (1996) make a useful distinction by referring to this as the *probable cause* LR; i.e., the LR where the suspect is arrested for reasons unconnected with his DNA profile. We have seen that, under certain assumptions, this LR reduces simply to the inverse of the match probability, $1/P$. Now consider that we have additional information: the suspect has been found as a result of a search of a database of N suspects. This particular suspect matched the crime genotype and the other $(N-1)$ did not. We have already seen how to deal with the event that the suspect and crime sample have the same genotype; let us use D to denote the additional information that the other members of the database did not. Then the LR in this case is

$$\mathrm{L}R \;=\; \frac{\Pr(E, D|H_p, I)}{\Pr(E, D|H_d, I)} \tag{9.2}$$

Let us refer to this as the *database search* LR; we need to see how it differs from the probable cause LR. First, let us expand it using the multiplication law for probability as follows:

$$\mathrm{L}R \;=\; \frac{\Pr(E|H_p, D, I)}{\Pr(E|H_d, D, I)} \frac{\Pr(D|H_p, I)}{\Pr(D|H_d, I)}$$

The first of these two ratios is very similar to the probable cause LR, Equation 9.1, save that the conditioning has been extended to include the information that none of the $(N-1)$ other suspects in the suspect's database has the crime profile G_C. The numerator is the probability of a match between suspect and crime stain given that the former left the latter: clearly the fact that there has been a database search does not affect this probability. The denominator is the match probability given the additional information that among $(N-1)$ other suspects there is no one with genotype G_C. It is clear that this extra information cannot increase the match probability; on the contrary, the fact that none of the other members of the database has that genotype increases confidence in its rarity and might tend to decrease the match probability. So this first ratio is approximately equal to (or slightly greater than) the probable cause LR of $1/P$. It follows that the database search LR is at least as large as the probable cause LR multiplied by the

ratio R:

$$R = \frac{\Pr(D|H_p, I)}{\Pr(D|H_d, I)} \tag{9.3}$$

Now the problem is clearly specified: it is necessary to determine whether R is smaller than or greater than one. Let us first consider the numerator, which is the answer to the question

> If this suspect is the person who left the crime stain, what is the probability that none of the other $(N-1)$ suspects would match the crime stain?

Note that the conditioning does not include the crime stain genotype, so the question is concerned with the general discriminating power of the profiling system. We now write $\psi_{(N-1)}$ for the probability that none of $(N-1)$ innocent people will have a genotype matching that of any unspecified crime stain. This is the numerator of R in Equation 9.3. Now consider the denominator, which is conditioned on the proposition that the suspect did not leave the crime stain. Now the question is

> If this suspect is not the person who left the crime stain, what is the probability that none of the other $(N-1)$ suspects would match the crime stain?

We have to allow for two possibilities. Given that the suspect did not leave the crime stain it is certain that someone else did. Either that other person is among the other $(N-1)$ suspects, and we denote this probability by ϕ, or he is not, with probability $(1-\phi)$. So the denominator is equal to ϕ times

$$\Pr(\text{No } \textit{match among the } (N-1) | \text{T}\textit{he offender is one of them})$$

plus $(1-\phi)$ times

$$\Pr(\text{No } \textit{match among the } (N-1) | \text{T}\textit{he offender is not one of them})$$

If this other person–the true offender–were among the $(N-1)$, then the chance that he would *not* match would be zero, so the product of the first pair of terms is zero. The second term of the second pair is just $\psi_{(N-1)}$, so the denominator is then $(1-\phi)\psi_{(N-1)}$. Combining the numerator and denominator shows that R is equal to $1/(1-\phi)$. We may feel that it is problematic to assign the probability ϕ, but this is not important. What is important is that the denominator is less than one, so R in Equation 9.3

is greater than one. Although it might not differ greatly from one, we have shown that the database search LR is, albeit slightly, greater than the probable cause LR. This is in direct contradiction with the NRC report. Note also that, despite our possible difficulty with assigning a numerical value to ϕ, it seems reasonable to infer that the larger is N, the greater should be ϕ, so the LR at Equation 9.2 *increases* with increasing N, rather than decreasing as does the NRC solution. In the extreme case in which N equals the entire population of the country, ϕ must be equal to 1 and the LR is infinite: this corresponds to the situation in which the suspect is the only person in the country who has the genotype of the crime stain. The logic copes with this in an entirely reasonable manner, whereas the NRC recommendation does not give a sensible answer.

We have seen that, at court, the prosecution proposition would be of the form

> H_p: The suspect left the crime stain.

which is naturally tested against the defense alternative of the form

> H_d: Some person other than the suspect left the crime stain.

The NRC recommendation comes from considering a defense alternative of the form

> H_d: Some person other than the N people in the database left the crime stain.

However, the other $(N-1)$ people in the database are not on trial.

If the NRC recommendation were adopted, then it would suggest a shrewd defense strategy arising from an example given by Balding and Donnelly (1996). Imagine that a suspect has been apprehended for reasons unconnected with the DNA database. In fact, he has never been in that database. He is found to match and the case is taken to court. Then defense counsel asks for a search to see if there any people on the DNA database with the same profile. If there are one or more database matches, the defense counsel can argue that the impact of the DNA evidence against the defendant is weakened. If there are no database matches the defense must argue that the court faces a situation corresponding to that considered by the NRC because there are a matching suspect and N nonmatching suspects. It follows that the match probability must be increased by multiplying it by N (actually by $N+1$ in this case because that is the total number profiled). Whatever the outcome of the database search, the defense case appears to be strengthened.

WRITTEN STATEMENTS AND REPORTS

Now we consider the manner in which the scientist should convey his results in a written report or statement. In Chapter 2 we laid out three principles for the interpretation of evidence, and these will guide the way in which we approach the task of writing the statement for this particular case.

Circumstances

The third principle in Chapter 2 was that the scientist must evaluate the evidence, not only under the competing propositions, but also under the conditioning of the non-DNA evidence, which we are referring to here as the "circumstances." If the interpretation depends on the circumstances then it naturally follows that the scientist should describe those circumstances in his statement. Of course, what we should say is *alleged circumstances* because the scientist will generally be presented only with the circumstances as they appear to the investigator. These could well change as the investigation and subsequent trial progress. Two objections could possibly be raised: first that the scientist's view of the circumstances might be regarded as hearsay, and second that the circumstances may change. The first is easy to deal with because the scientist is not testifying to the validity of those circumstances: he is only repeating what he believes to be evidence to be presented by other witnesses. The second is actually an argument *in favor* of stating the perceived circumstances. It is essential that readers of the statement understand that the interpretation has taken place in a framework of circumstances. Furthermore, if the circumstances *do* change, then the scientist will need to review the interpretation and a sentence to this effect should, ideally, be included in the statement. It is a matter of judgment to decide which aspects of the circumstances need to be stated. Only those that are relevant to the interpretation should be included. In the present example these would be

- The alleged offense occurred in Gotham City.
- The offender was described as Caucasian and spoke with a local accent.
- The suspect has no close blood relatives in the city who would be considered suspects.

It is because of these features of the circumstances that the scientist has considered it most fitting to use a Caucasian database. The fact that his database was collected from Gotham City increases his confidence in

its relevance because the offender appears to be local. It is because of the absence of blood relatives that he has calculated a match probability for an unknown person unrelated to the suspect, and his decision to work on the basis that the suspect and offender belong to the same subpopulation seems suitably cautious. The other aspects of the circumstances, such as the height, build, hair color, and age of the suspect and offender, while highly relevant to the deliberations of a court, do not appear relevant to the interpretation of the DNA evidence and can therefore be omitted from the statement.

Alternatives

We have seen that the first principle of interpretation states that it is not meaningful to address the uncertainty with regard to the truth of a proposition without considering at least one alternative proposition. In the present case there is one proposition that, at first sight at least, appears clearly defined. It is that of the investigator: Mr. S. raped Ms. V. It may be tempting for the scientist to take such an alternative as the first proposition for explaining the evidence, but a moment's reflection suggests that this might not be a wise course because the DNA profiling evidence cannot shed light on whether or not Ms. V. was actually *raped*: Mr. S. might later confirm that he and Ms. V. did have sexual intercourse but allege that it was with her full consent. In such an eventuality, the DNA evidence provides no assistance for weighing the prosecution and defense propositions against each other. Further thought suggests a proposition that is a stage or two removed from what prosecution will set out to prove. In this case it would seem sensible to suggest

> H_p: The semen on the vaginal swab came from Mr. S.

The problem that we next face appears peculiar to the legal field because the alternative proposition should reflect the position of the defense but, of course, the defense in most jurisdictions are under no obligation to put forward any explanation for the evidence. The scientist must therefore anticipate a defense proposition and might decide to address something of the form: The semen on the vaginal swab did not come from Mr. S. Once again, reflection suggests that this may not be a helpful way of expressing the alternative. Let us recall what the scientist did. He calculated a match probability for an unknown person unrelated to Mr. S., and this must surely show that the alternative proposition being addressed is

> H_d: The semen came from some unknown Caucasian man who was unrelated to Mr. S.

The scientist could, if he wished, make this still more specific to state that the unknown man came from the same subpopulation as Mr. S., but such a refinement is probably not needed and it may obscure clarity. The alternative propositions to be made clear in the statement are thus

> H_p: The semen on the vaginal swab came from Mr. S.
> H_d: The semen came from some unknown Caucasian man who was unrelated to Mr. S.

The scientist should also state his willingness to address other alternative propositions if necessary.

Evaluation

We have said that, given the defense proposition, the match probability is one in a million, and this is one way of expressing the strength of the evidence. The other way is to quote the LR of one million, because in this case the probability of the evidence given the prosecution proposition is one. In a case such as this, where the DNA evidence is very simple, there is no strong reason for choosing between the two methods. But this is not necessarily true when it comes to presenting the evidence in court, as we shall see shortly. Certainly, if the pattern of evidence is more complicated, such as when the crime sample is a mixture, then there is no sensible alternative to the LR, so we would favor always quoting the evaluation in that form, if only for consistency. In the present case that would mean saying something like, "The evidence is a million times more probable given the first alternative." We will discuss comprehensibility later when we talk about presenting evidence at court, but for the moment we remark that we may feel that we could do more to convey the weight of evidence to nonscientists than merely quote a number. We know that the readers of the statement will not be practiced at dealing with numbers. One way of helping with this is to augment the statement with some kind of verbal supplement. The British Forensic Science Service addressed all the issues of communicating by means of statements in the late 1980s, and one of the products of that review was a move toward standardization centered on a small subset of words to be used whatever the evidence type.

Verbal Conventions

It has been a widespread practice in the forensic science community to rely heavily on the phrases "could have" and "consistent with." In the present case, for example, we might report that the semen on the vaginal swab could

have come from Mr. S. At best this is no more than a statement of the obvious: at worst it is a surrogate for a scientific interpretation when, in fact, it contains no element of interpretation. The same goes for "consistent with." Furthermore, such phrases could be seen as indicating a prosecution bias unless the other side of the coin is presented. In this case that would mean saying that the semen could have come either from Mr. S. or from anyone else of the same genotype. Such sentences merely fill space and convey no useful interpretation. It is our experience that the phrases "could have" and "consistent with" serve only to obscure clarity of thought, and we recommend against their use in forensic statements. The likelihood ratio suggests a simple verbal convention based on the use of the word "supports." Statisticians (Edwards 1992) have defined SUPPORT as the natural logarithm of a likelihood ratio in order to speak of the support provided for one hypothesis against another by some data. We will not adopt such a formal definition of support, but we do like the word in this context. In particular, if the likelihood ratio exceeds one in the kind of example we have been considering, we will say that the evidence supports the prosecution proposition. If it is less than one, then it supports the defense proposition. It is natural, in any given case, for the court to ask "How strong is the support?" and the convention then comes down to a range of qualifiers that relate to broad ranges of the likelihood ratio. A possible convention might be of the form

Likelihood ratio	Verbal equivalent
1 to 10	Limited support
10 to 100	Moderate support
100 to 1000	Strong support
more than 1000	Very strong support

We recognize that there is much to debate. Certainly, these verbal equivalents cannot be seen to be cast in stone, and if a number has been calculated, then the verbal statement can be added only as an aid for greater understanding. For non-DNA evidence, where numerical likelihood ratios are not calculated, the restricted range of language serves to remove one source of variability from discussion of the issues. It seems to us that language becomes inadequate for likelihood ratios in the tens and hundreds of thousands or higher, and we do not have a category above "very strong."

DNA EVIDENCE AT COURT

So far, we have explained what we see to be the elements of balanced reporting of DNA cases. If the case is taken to court, then the scientist meets

the next challenge of his profession–that of interacting with members of the legal profession who are generally not familiar with the principles of interpretation. This can be an excruciatingly difficult time for the scientist, particularly if the time available for consultation with the lawyer who is to lead the questioning is scant, or worse, if the scientist is faced with a lawyer who is not prepared to gain any sort of understanding of the principles. DNA profiling presents new challenges to courts because it is the first type of transfer evidence for which the weight of evidence is encapsulated in a number. (We are excluding paternity testing here.) Previously, courts have readily accepted expert opinion phrased in terms such as

> In my opinion this mark was made by the left forefinger of the defendant.
> I consider it highly probable that the defendant wrote this handwriting.

The acceptance of such statements bypassed any statistical considerations, and far more is known about the statistics of DNA profiling than about the statistics of fingerprint minutiae or of handwriting characteristics. The notion that each corresponding fingerprint minutia adds weight to the evidence has long been accepted with little justification based on survey data. This is in remarkable contrast to the battles that have been waged over the use of the "product rule" to combine information from the constituent alleles in a DNA profile. The new emphasis on quantification of evidence has led to perplexing courtroom decisions and it has been difficult for forensic scientists to pursue logical arguments with confidence, in the face of a lack of understanding from all sides.

We will return later to the issue of individualization, but now we continue with our Gotham City example and imagine that the decision has been made to prosecute Mr. S. The scientist now has the job of presenting the evidence in court. If the statement has been written along sound lines then he starts from a strong position for answering challenges in court. However, no matter how strong that position is, he may still face questions that are difficult, not because they are especially probing or even designed specifically to confound him. It is paradoxical that the questions that stem from the most well meaning attempts by counsel to create greater enlightenment can, in fact, have the contrary effect of actually making things more confused. Such questions can lead to fallacious lines of reasoning. We now discuss a few of these and illustrate them with examples from a few cases heard in the British courts.

The Transposed Conditional

In Chapter 2 we met the first and most pervasive kind of distortion that the scientist is likely to meet: the transposed conditional. Although it was dubbed "the prosecutor's fallacy" by Thompson and Schumann (1987) it is an error made by prosecutors, defenders, and judges. Furthermore, even if the scientist has made every effort to get things right, he will find that he has been misquoted as giving a probability for the proposition that the crime sample came from someone other than the defendant. For example, the Associated Press on November 14, 1996, reported

> The chances are at least 1-in-170 million that anybody else's DNA besides Simpson's could be contained in a blood drop found near the bodies of Nicole Brown Simpson and Ronald Goldman, testified Robin Cotton, director of Cellmark Diagnostics in Maryland.

The reporter was confused between the meaning of "at least" and "at most" but, more importantly, misquoted what Dr. Cotton said. Furthermore, on November 21, 1997, a story in the scientific journal Science began with

> Even in O.J. Simpson's trial, prosecutors could only say that the odds were billions to one that blood found at the scene was not O.J.'s.

Let us look at how the error could manifest itself in the Gotham City example. The scientist might be asked to express the weight of evidence by means of the match probability, in which case he might say something of the form

> The probability of a match if the semen came from someone other than Mr. S. is one in a million.

Then, in an effort to make things clearer for the jury, counsel might lead with a question of this form:

> You mean that there is only a one in a million chance that the semen came from someone other than Mr. S.?

This is but a short step from inferring that the odds are a million to one on that the semen came from Mr. S., and the error is well and truly compounded. The scientist in this example has started on a sound footing by talking about the probability of the evidence. Furthermore, he has made the conditioning very clear by the use of the word "if." That starting point

provides a good basis for keeping the questioning on track. If the evidence is not stated clearly then a slip to the transposed conditional is even easier. Here is an example from the UK trial of Andrew Deen (*R. v. Deen*, *The Times*, 10 January 1994):

> Q: So the likelihood of this being any other man than Andrew Deen is one in three million?
> A: In three million, yes.

Counsel had clearly committed the prosecutor's fallacy and the scientist lost an opportunity to correct him. Here is an extract from the trial of Alan Doheny (*R. v. Doheny, R. v. G. Adams* [1997] 1 Cr App Rep 369):

> Q: What is the combination, taking all those [RFLP bands] into account?
> A: Taking them all into account, I calculated the chance of finding all those bands and the conventional blood groups to be about one in 40 million.
> Q: The likelihood of it being anybody other than Alan Doheny?
> A: Is about one in 40 million.

Here, of course, the scientist clearly acceded to the prosecutor's fallacy and this was exposed on appeal. It might be thought that presenting the evidence in the form of a likelihood ratio provides a solution to the problem of avoiding the transposed conditional. Alas, it is not so. A statement of the form "The evidence is a million times more probable if the semen came from Mr. S. than if it came from someone else," can be turned, incorrectly, to, "It is a million times more probable that the semen came from Mr. S." We have to accept that the trap of the transposed conditional always awaits the unwary. There are a few precepts to help minimize the danger. First, start with a statement founded firmly on the principles we have stated. Second, never offer an opinion for a proposition given the evidence: always speak about the probability of the evidence given the proposition. Third, always make the conditioning clear in probability statements by the use of the word "if" or "given."

There is a point of view (expressed to one of the authors on one occasion by a learned judge!) that if the difference between the logically correct and incorrect sentences is apparently so subtle then perhaps the fallacy doesn't matter. No matter how careful the scientist, lawyer, and judge may be, for all we know jurors will mentally transpose the conditional anyway. Here is an interesting extract from the Appeal Court judgment in *R. v. Doheny* and *R. v. G. Adams*, when they argued that the prosecutor's fallacy had not necessarily misled the jury:

> Given [the match probability of] one in 40 million, we have no doubt that the Jury would have reached the same verdict if directed in this way. The more remote the random occurrence ratio, the less significant will be the adoption of the "Prosecutor's fallacy", until the point is reached where that fallacy does not significantly misrepresent the import of the DNA evidence. Such was the position of the figures advanced by [the forensic scientist].

Nevertheless, the Appeal Court clearly ruled that the fallacy should be avoided:

> The scientist should not be asked his opinion on the likelihood that it was the Defendant who left the crime stain, nor when giving evidence should he use terminology which may lead the Jury to believe that he is expressing such an opinion.

Whatever we may think about the apparent subtlety of the fallacy and its impact on the jury, it *is* a fallacy and a forensic scientist cannot contribute to its use. The final argument is that, if committed, it will always provide grounds for appeal. The appeals of both Deen and Doheny were allowed, in each case partly because of the prosecutor's fallacy.

Fallacy of the Misapplied Expectation

Recall that in our Gotham City example the scientist had given a match probability of one in a million, and recall also that the population of the city is approximately two million. Here is a line of reasoning: there are roughly one million males in the city; the chance that any one would have the crime profile is one in a million; therefore we can expect to find one man in the city with that profile; Mr. S. must be that man. At an intuitive level this may appear a compelling argument, yet it has two serious flaws which we will shortly discuss. It is a line that has actually been used in court, however, and we quote from *R. v. G. Adams:*

> Q: Is it possible that the semen could have come from a different person from the person who provided the blood sample?
> A: It is possible but it is so unlikely as to really not be credible ... I can estimate the chances of this semen having come from a man other than the provider of the blood sample. I can work out the chance as being less than one in 27 million.
> Q: So it is really a very high degree of probability indeed that

> the semen stain came from the same person who provided the blood sample?
> A: Yes. You really have to consider the size of the group of individuals who could possibly be the source of this semen. Now, there are probably only 27 million male people in the whole of the United Kingdom so a figure of one in 27 million does tend to imply that it is extremely likely there is only really one man in the whole of the UK who has this DNA profile.

The first answer is clearly a transposed conditional. The third is the same fallacious line as we have put forward for our Gotham City example. The judge in *R. v. G. Adams* appeared to pick this last point in his summing up:

> That means ... not less than one in 27 million people ... I should not think there were more than 27 million males in the United Kingdom, which means that it is unique.

As an aside, note that the judge made a similar mistake to the one we saw in the report of Dr. Cotton's evidence earlier in this Chapter: He should have said either "one in not less than 27 million people" or simply left the "not" out from his sentence. This is a very common error when frequencies are quoted but it does not arise when the evidence is given in the form of a likelihood ratio.

Returning to the Gotham City example, there are two flaws in the argument. The first lies in a misunderstanding of the nature of an expectation, which we introduced in Chapter 3. It is a trivial calculation to multiply one in a million by a million to get the answer one–but this is an expected number. It does not mean that there must be one man with probability one. The Poisson distribution, which we met in Chapter 3, may be used here. In this case the Poisson parameter λ is one and we can calculate the probability of 0, 1, 2, 3 ... individuals, using the formula for the Poisson pdf, as 0.368, 0.368, 0.184, 0.061 ... respectively: readers may wish to confirm these probabilities using Equation 3.2. So, although the expectation is one, far from there being certainty of there being one and only one, there is a 0.264 probability of there being more than one. The second flaw is more serious. Recall that we have calculated the match probability conditioned on the knowledge of the defendant's genotype. It is the probability that *a person other than the suspect* would have that genotype. If we want to talk about numbers of men we should be asking, given that we know that the defendant has that genotype, how many *other* men do we expect to find. Ignoring for the moment the issue of subpopulations and evolutionary relatedness, we return

to the Poisson. The total number of other men in Gotham City is reduced by one, because we have set the defendant aside, but this obviously has no discernible effect on the calculation: the Poisson parameter is one, as before. Then the probabilities of 0, 1, 2, 3 ... *other* men are as listed in the previous paragraph. So the probability that there is at least one other man in Gotham City is approximately $0.368+0.184+0.061+\ldots = 1-0.368 = 0.632$.

The position with regard to the defendant is more easily seen if we use a Bayesian argument. If we consider it reasonable to consider all million males in the city as equally likely to have been the offender, in the absence of other evidence, the prior odds in favor of Mr. S. being the offender are a million to one against. The likelihood ratio for the DNA evidence is one million, so the posterior odds are one, or *evens*. Based on the uniform prior and the DNA evidence there is a 0.5 probability that Mr. S. was the assailant. The Bayesian arithmetic is just the same for the numbers cited in the G. Adams trial. Of course, even this is a weak method of reasoning. The idea that all males in the city should be regarded as equally likely suspects is quite unrealistic, as we discussed in Chapter 3. We return to such lines of argument later.

Concluding "It's Him."

It is almost inevitable, in any discussion of DNA evidence, that the question will be asked "Is it as good as fingerprints?" Courts throughout the world have been accustomed for decades to experts giving opinions of the form "This control print and this mark from the crime scene were made by the same person." It is not surprising that occasions have arisen where experts were prepared to give opinions of similar force with regard to DNA comparisons. For example, from *R. v. Deen*:

> Q: On the figures which you have established according to your research ... what is your conclusion?
> A: My conclusion is that the semen has originated from Andrew Deen.

From the trial *R. v. Doheny*:

> Q: You deal habitually with these things, the jury have to say, on the evidence, whether they are satisfied beyond doubt that it is he. You have done the analysis, are you sure that it is he?
> A: Yes.

And from *R. v. G. Adams*:

> Q: Is it possible that the semen could have come from a different person from the person who provided the blood sample?
> A: It is possible but it is so unlikely as to really not be credible.

The ruling in *R. v. Doheny* quoted in the section on the transposed conditional now makes it clear that, in UK courts, scientists giving evidence on DNA are not to be asked to express an opinion as to whether a trace from a crime scene came from the defendant. DNA evidence alone, unlike all other kinds of transfer evidence, is peculiar in this regard. It might appear paradoxical that, in the single kind of transfer evidence for which meaningful statistical assessments can be made of the weight of evidence, the scientist is not to be permitted to express an opinion of origin. Vastly more is now known about DNA statistics than about handwriting, fingerprints, toolmarks, footwear, and so on, yet that evidence type alone is the one in which the scientist is not to express an opinion of origin. Later in the chapter we look at the reasons this should be so.

Size of Population Subgroups

We have seen that the Appeal Court in *R. v. Doheny* and *R. v. G. Adams* ruled that the expert was not to express an opinion on whether or not a given crime sample came from the defendant. The question then is: If the expert is not to give that guidance, how is the jury to reach a decision? The Appeal Court made some suggestions about how they may be assisted by the scientist, as follows:

> He will properly explain to the Jury the nature of the match ... between the DNA in the crime stain and the DNA in the blood sample taken from the Defendant. He will properly, on the basis of empirical statistical data, give the Jury the ... frequency with which the matching DNA characteristics are likely to be found in the population at large. Provided he has the necessary data, and the statistical expertise, it may be appropriate for him then to say how many people with the matching characteristics are likely to be found in the United Kingdom–or perhaps in a more limited relevant subgroup, such as, for instance, the Caucasian sexually active males in the Manchester area. This will often be the limit of the evidence which he can properly and usefully give. It will then be for the Jury to decide, having regard to all the relevant evidence, whether they are sure that it was the Defendant who left the crime stain, or whether it is possible that it was left by someone else with the same matching DNA characteristics.

The suggestions of the court have seductive intuitive appeal but they offer many traps to the unwary. If the match probability in the case in question is relatively large, then answering a question of the kind, "How many people with the matching characteristics are likely to be found in the United Kingdom?" is simple. For example, if the match probability is one in a million, then one would expect to find around 260 matching people in the US population of 260 million. It is by no means clear how this helps the jury. As we have said before, it is impossible to visualize a case in which all of the members of the populace can be viewed as equally likely suspects for the crime. The Appeal Court appeared to anticipate this problem by suggesting that it might be possible to narrow the pool of suspects–in their example to the "Caucasian sexually active males in the Manchester area." What does this mean? How do we define "sexually active"? What is "the Manchester area"? Is it just the city of Manchester, or should it include all the towns in the Greater Manchester area? How much of Cheshire and Merseyside should be included? Let us imagine, purely for illustration, that by some process or other the prosecution and defense in a trial decide that the number of sexually active males in the suspect population is 500,000. Then a match probability of one in a million suggests that the number of men with matching profiles in the suspect population is expected to be one-half. How is that to be explained to the jury? If the suspect population were only 10,000 men, then the expected number of other men with the same genotype is 1/100. In such a situation it might be considered permissible for the scientist to offer the opinion that he considered it unlikely that another person with that genotype would be found in that group. But this approach, too, is not without its problems, and the scientist may be safer here if he leaves the field to a qualified statistician.

However it is done, the line suggested by the Appeal Court means that the prosecution in a DNA trial will want to reduce the numbers in the suspect population. Let us look at how that might proceed in our Gotham City example. We saw in the introduction that Ms. V. described her attacker as Caucasian, around 1.8 m tall, with dark hair, medium build, in the age range 25 to 30 years. She said also that he spoke with a local accent. We assume that this evidence is put to the jury as part of Ms. V.'s evidence. In turn, the jury will learn that Mr. S. is indeed a local Gotham City resident, 1.7 m tall, with dark hair, medium build, aged 35 years.

The forensic scientist would start with his calculation that, based on the DNA profile, he would expect there to be one other person in Gotham City with the same genotype. Prosecuting counsel may then invite him to consider the consequences of reducing the size of the suspect population

using the other aspects of the evidence. Ms. V. described her attacker as 1.8 m tall: the prosecution might ask the scientist how this reduces the size of the suspect population. It is doubtful if any forensic scientist would be prepared to enter this kind of exercise, but even if our expert were so inclined, what is to be done? Do we consider men of exactly 1.8 m? Does the prosecution conveniently widen the height window to include men of 1.7 m, that being the height of the defendant? What about using age to delimit the suspect population? Presumably there are demographic tables for Gotham City that will tell us the number of men in the city aged 25 to 30 years. Again the prosecution might make a concession of extending the age window to 25 to 35 years to include the defendant. This would have possibly been acceptable if the victim had said that her attacker was in the age range 25 to 35 but that is not the case–if Mr. S. is the attacker then she *got his age wrong.* Neither does her evidence in relation to the height of her attacker fit very well with Mr. S. Forensic scientists should be wary of being drawn into such murky waters. The "shaving down the population" approach suggested by the Appeal Court ruling cannot work if there is conflict in the non-DNA evidence.

The procedure of addressing the other evidence becomes much more obvious if one looks at it from the perspective of the principles of interpretation described in Chapter 2. The questions to be addressed are of the following kind:

- What is the probability that Ms. V. would say that the offender was around 1.8 m tall if the offender was Mr. S?
- What is the probability that Ms. V. would say that the offender was around 1.8 m tall if the offender was some unknown man?
- What is the probability that Ms. V. would say that the offender was in the age range 25 to 30 if the offender was Mr. S?
- What is the probability that Ms. V. would say that the offender was in the age range 25 to 30 if the offender was some unknown man?

We are not suggesting that the forensic scientist should attempt to lead the court through such a list of questions. Even though this is a distinct improvement on the prosecution-led reduction in population size, there are potential problems. The two sets of questions above, for example, relate to attributes that are not independent of each other, and care is needed in framing them rather more precisely than we have done. There has been a case, nevertheless, in the British courts where such an exercise was carried through and we will now discuss it.

Bayesian Reasoning for Nonscientific Evidence

In 1991 a Ms. M. was walking home alone in the early hours of the morning in Hemel Hempstead, England. A male stranger approached her and asked the time. She saw his face for a few seconds before looking at her watch. He then attacked her and raped her from behind. She later told police that the attacker was Caucasian, clean-shaven, with a local accent, aged between 20 and 25 years.

No suspect for the offense was found until 1994, when Dennis John Adams came to notice in connection with another incident, and his DNA profile was found to match that from the semen recovered from the rape of 1991. As a result of this, he participated in an identity parade, at which Ms. M. failed to pick him out. Furthermore, at subsequent committal proceedings she said that Mr. Adams did not look like the man who attacked her. His age at that time was 37, but she said that he appeared to her to be around 40 to 42–appreciably older than the description of her attacker.

At trial, prosecution produced the DNA evidence which was associated with a likelihood ratio of 200 million (this was disputed by defense, but for the sake of this discussion it will suffice to use this figure). With the exception of the locality–Mr. Adams lived near to the scene of the rape–all of the other evidence favored Mr. Adams' innocence. In addition to the conflicting identification evidence, he gave an alibi for the night of the attack, which he said he spent with his girlfriend.

This is a case that illustrates the weaknesses of the guidance given by the Appeal Court in *R. v. Doheny* and *R. v. G. Adams*. If we accept that the match probability was 1 in 200 million, then the expected number of adult males in the UK who would have the same profile is about 1/10: it is not at all clear how this would assist a jury. What if we now attempt to whittle down the size of the suspect population using the victim's description? Consider the age: a prosecutor might ask for the number of males aged 20 to 25 years. But this would be misleading, because the defendant himself did not fall into this age group. So it might be argued that the interval should be widened to include all males up to the age of 40, say. But even this would be inadequate, because it takes no account of the fact that the defendant's age does not agree with the victim's description.

If we seek a logical evaluation of the evidence in this case, as we have already mentioned in our illustrative Gotham City example, it is necessary to apply Bayesian reasoning to all the nonscientific evidence and this is exactly how defense counsel argued at the trial. A statistician (P. Donnelly) was called to explain how this could be done by the jury. Of course, we do not know how the jury actually reasoned during their deliberations, but the

outcome was that Mr. Adams was found guilty and sentenced to a prison term.

Defense were granted leave to appeal, and the Appeal Court overturned the conviction and ordered a retrial on the grounds that the trial judge had given insufficient guidance to the jury on other methods for evaluating the evidence. The appeal had been sought on several grounds, but it was allowed for reasons associated with the use of Bayes' theorem. Here are extracts from the judgment (*R. v. D. Adams* [1996] 2 Cr App Rep 467)

> It seems to us that the difficulties which arise in the present case stem from the fact that, at trial, the defense were permitted to lead before the jury evidence of the Bayes theorem ... we have very grave doubt as to whether that evidence was properly admissible, because it trespasses on an area peculiarly and exclusively within the province of the jury, namely the way in which they evaluate the relationship between one piece of evidence and another.

Several reasons were given, including

> ... the attempt to determine guilt or innocence on the basis of a mathematical formula, applied to each separate piece of evidence, is simply inappropriate to the jury's task. Jurors evaluate evidence and reach a conclusion not by means of a formula, mathematical or otherwise, but by the joint application of their individual common sense and knowledge of the world to the evidence before them.

The strength of the Appeal Court views is summarized by the following

> Quite apart from these general objections, as the present case graphically demonstrates, to introduce Bayes theorem, or any similar method, into a criminal trial plunges the jury into inappropriate and unnecessary realms of theory and complexity deflecting them from their proper task.

And, in conclusion

> If, as seems entirely possible, the jury abandoned the struggle to understand and apply Bayes, they were left by the summing-up with no other sufficient guidance as to how to evaluate the prosecution case (based as it was entirely on the DNA evidence), in the light of the other non-DNA evidence in the case. This means that their verdict cannot be regarded as safe.

At the retrial, defense argued in a manner similar to its line at the first trial. Since the DNA evidence was presented numerically, reasoned defense counsel, the only way in which the jury could effectively consider the overall weight of evidence was to consider the non-DNA evidence numerically. The appropriate way to do this was by means of Bayes' theorem. The view of the statisticians who were advising prosecution was that this was too difficult an exercise for the jury, but if it were to be done, the prosecution offered its own experts to work with the defense expert in the preparation of a questionnaire for the jury to use. This was duly done. There were over 20 questions, directed at assessing the value of nonscientific aspects of the evidence. Here is an example, which is a pair of questions relating to the victim's description of her attacker:

> Bearing in mind that Ms. M. said that Mr. Adams appeared to be in his early forties what is the probability that she would say that her assailant was in his early twenties if Mr. Adams were indeed the assailant? What percentage of assailants in cases of this nature would be described as being in their early twenties?

The questions were designated by letters of the alphabet and questionnaire ended in a formula for combining the answers. The use of the questionnaire was explained in detail from the witness box by the defense statistician. Mr. Adams was again convicted and, once more, appealed. The following are extracts from the judgment of the court that considered the second appeal (*R. v. D. Adams*, 16 October 1997, CA 96/6927/Z5):

> If ... the jury concluded that they did accept the DNA evidence wholly or in part called by the Crown, then they would have to ask themselves whether they were satisfied that only X white European men in the United Kingdom would have a DNA profile matching that of the rapist who left the crime stain. It would be a matter for the jury, having heard the evidence, to give a value to X. They would then have to ask themselves whether they were satisfied that the defendant in question was one of those men. They would then go on to ask themselves whether they were satisfied that the defendant was the man who left the crime stain, bearing in mind on the facts of this case the obvious discrepancies between the victim's description of her assailant and the appearance of the appellant, ... Of course, it is a matter for the jury how they set about their task, and it is no part of this court's function to prescribe the course which their deliberations should take. But consideration of this case along the lines

> indicated would in our judgment reflect a normal course for a properly instructed jury to adopt. It is the sort of task which juries perform every day ... as they are sworn to do.
>
> We are very clearly of opinion that in cases such as this, lacking special features absent here, expert evidence should not be admitted to induce juries to attach mathematical values to probabilities arising from non-scientific evidence adduced at the trial.

The appeal was dismissed.

It might be thought that the rulings were against the Bayesian view of scientific evidence. This is not correct: the judges were ruling against its application for the evaluation of *nonscientific* evidence. This, in fact, corresponds to the advice given to the prosecution by its own experts in both trials. But the judgment is, ultimately, unsatisfactory. The jury faced a difficult task in this case–powerful evidence supporting Mr. Adams' involvement, yet persuasive evidence from the victim herself supporting his innocence. The judgment clearly states that the jury should not be expected to tackle this problem logically. However, they are expected to assign a value X to the number of men in the country who would match the crime profile, though we have already seen that this line of reason can be problematic for a scientist, let alone a lay person. If the jury decided that $X = 1/10$, for example, how is this to be related to the conflict between victim's description of her attacker and the appearance of the defendant? This is hardly "the sort of task which juries perform every day." The root problem is that the existing legal system exerts powerful forces against carrying through the most appropriate procedure effectively. Notwithstanding the efforts put into the design of the questionnaire and its explanation to the jury, it must still have remained a puzzling exercise to them. However, if sufficient time and resources had been forthcoming, then a computer-based dialogue could have been produced. If free dialogue between the expert and jury had been possible, then the system could have been tailored to the intellectual capabilities of each individual jury member. Although such a solution is well within the bounds of today's technology, the procedural difficulties would be immense because of the nature of the established institution.

INDIVIDUALIZATION AND IDENTIFICATION

It will be useful at this point to look at some general issues relating to the process of identification and then to consider the process of fingerprint comparison.

Individualization

In the forensic sphere, the words *identity* and *identical* tend to be misused. Examiners sometimes give opinions of the kind "these two marks are identical." This is not correct because any entity can only be identical with itself. Two marks, whether they are from the same finger, from the same item of footwear or made by the same tool, cannot be *identical* and indeed they will inevitably be *different* in detail. Two different entities cannot be identical to each other because they are each unique. This applies not only to so-called "identical twins" but also to all of the grains of sand on a beach. Note also that a DNA profile is a manifestation of a complex biological/physical/chemical process and two DNA profiles cannot be identical to each other, even if they have come from the same person. The fact that we choose to summarize each profile by a set of numbers and that two profiles have the same sets of numbers merely means that they are indistinguishable from each other using the measuring system that we have chosen.

The issue for the forensic scientist is not "Is this profile unique?" (it is) or "Are these two things identical?"(they can't be) but "Is there sufficient evidence to demonstrate that they originate from the identical source?" We notice that it is widespread practice in the forensic field to refer to the process that leads to the answer "yes" to this question as *identification.* Kirk (1963) pointed out that the word *individualization* was more appropriate in this regard; indeed, he defined criminalistics as the "science of individualization." Nevertheless, we must bow to what has become general usage–certainly in the fingerprints field–and refer to a categorical opinion of identity of source as an "identification."

DNA versus Fingerprints

Earlier we raised the question "Is a DNA profile as good as a fingerprint?" It is important that we should understand a fundamental difference between the processes of inference that are pursued in the two fields, which was concisely explained by Stoney (1991):

> Fingerprint comparisons have the colloquial specificity of absolute identification, but a completely different [compared to DNA profiling] philosophy for achieving it. Although the study of fingerprint variation is founded on scientific observations, the process of comparison and the conclusion of absolute identity is explicitly a subjective process. The conclusions are accepted and supported as subjective; very convincing, undoubtedly valid, but subjective. In fingerprint comparisons, the examiner notes

> the details in the patterns of ridges. Beginning with a reference point in one pattern, a corresponding point in a second pattern is sought. From this initial point the examiner then seeks neighboring details that correspond in their form, position and orientation. These features have an extreme variability that is readily appreciated intuitively, and which becomes objectively obvious upon detailed study. When more and more corresponding features are found between two patterns, scientist and lay person alike become subjectively certain that the patterns could not possibly be duplicated by chance.
>
> What has happened here is somewhat analogous to a leap of faith. It is a jump, an extrapolation, based on the observation of highly variable traits among a few characteristics, and then considering the case of many characteristics. Duplication is inconceivable to the rational mind and we conclude that there is absolute identity. This leap, or extrapolation, occurs (in fingerprinting) without any statistical foundation, even for the initial process where the first few ridge details are compared. A contrast with our DNA individualization process is important because we in no way approach DNA evidential interpretation in the same way we approach fingerprints. We hold fingerprint specificity and individuality up as our ideal, yet this ideal is achieved (and can only be achieved) through a subjective process that we patently reject when applied to DNA. With DNA typing, as in conventional serological typing, we view our increasing evidential value as a step-wise process. We detect a series of traits, each one of which is, to some degree, rare. This leads to the inference of smaller and smaller joint probabilities and a conclusion that the combined type would be very very rare.

Stoney contrasts the undisputed subjectivity of a fingerprint comparison with what he sees to be the objectivity of a DNA statistic. Yet we have seen that this objectivity is itself an illusion because it exists only within a framework of assumptions. In the individual case it is for the scientist to judge the validity of those assumptions and to carry out whatever calculations he considers necessary given the case circumstances. In the wake of the Doheny/Adams appeal ruling in the UK, there has been a tendency for courts to seek a "statistical probability" or a "mathematical probability" in the mistaken belief that such numbers exist independently of human judgment.

We should be in no doubt about the degree of certainty implicit in a fingerprint identification. The expert is, in effect, saying "I am certain that this latent mark and this control print were made by the same person and *no amount of contrary evidence will shake my certainty.*" Or, to look at this from a Bayesian perspective, no matter how small the prior odds are, the likelihood ratio is so large that the posterior odds approach infinity. Stoney sees that a fingerprint identification is based on a "leap of faith," and he is quite correct to conclude that such a leap of faith has nothing to do with scientific principles. It is that leap of faith that characterizes the essence of a conclusion of identity of source and, as he points out, that is a fundamental difference between fingerprint evidence and DNA evidence. Stoney's "leap of faith" is equivalent to attaining an infinite likelihood ratio: this kind of belief cannot derive from any *scientific* process.

Ultimately, it must always be the jurors, or other triers of fact, whose belief in the proposition of an identical source that matters. The question is about the role that the scientist plays in determining that state of belief in the juror's mind. With conventional evidence types it has long been accepted in the courts that it is right and proper for the scientist to give his view on the proposition of an identical source and then it is a matter for the juror to decide on his confidence in the expert's judgment. However, the judgment in *R. v. Doheny* and *R. v. G. Adams* means that, for British scientists at least, that must not be done with DNA evidence.

So, in considering the question "Is it as good as a fingerprint?" we must recognize that a fingerprint identification is based on a process that is quite different in nature from that which we follow in interpreting a DNA match. The fingerprint identification means that the expert has reached a characteristic mental state of complete certainty based on the skilled and complex comparison that he has made. No juror is competent to attempt that comparison. As Stoney says, the expert does not *prove* individuality, he becomes mentally convinced by it. The issue is only proved when the court decides that he is competent to give that opinion and the jury decides that he can be believed.

With DNA, on the other hand, once the genotypes of the crime profile and the suspect have been determined, the comparison is trivial–any juror can see whether or not they are the same. Whereas the fingerprint expert does not consciously dissociate the two components, numerator and denominator, of the likelihood ratio, with a DNA match we are generally happy with the notion that the numerator is one and the assessment of the weight of the evidence comes down to considering the magnitude of the denominator. Once we assign a number to the denominator then we must recognize

that we have given the court something that they may choose to work with without our assistance. Certainly, the idea that the scientist has some particular power to take that number and take a step equivalent to the Stoney "leap of faith" is misconceived. If he really wishes to emulate the fingerprint expert, he must say "that match probability is so small that no amount of contrary evidence will shake me from the opinion that the crime sample was left by the defendant."

The expression "DNA fingerprinting" fosters an unrealistic impression of the technology, and it should not be encouraged in forensic circles.

Independence Across Loci

There is a key statement in the 1996 NRC report:

> We foresee a time when each person can be identified uniquely (except for identical twins).

The report contains language to the effect that, when DNA profiles match at a large number of loci, it is not reasonable to believe that they come from different people. This is based on our understanding that each person is genetically unique, identical twins excepted. The NRC statement reflects the widespread view that individualization through DNA profiling is a matter of testing at a sufficient number of loci. This is an understandable position to take: it appears to be inarguable that the more matching loci, the better the evidence. But how do we combine likelihood ratios from the different loci? Clearly, we would prefer to multiply them and justify this by an independence assumption. Providing the likelihood ratios are moderate enough that matches can be found in a database, then we can investigate the robustness of the assumption by suitable experiments based on between-person comparisons as we described in Chapter 5. As a rough guide, it seems reasonable that a likelihood ratio of, say, one million can be presented credibly if the scientist can quote a between-person experiment based on at least a million comparisons. The experiments of Lambert et al. (1995) and Evett et al. (1996) describe millions of comparisons, the former based on four-locus RFLP data and the latter on four-locus STR data. Larger experiments on RFLP data were conducted by Risch and Devlin (1992).

As we test more and more loci, we find larger and larger likelihood ratios for matching profiles, and we face two problems. First the credibility problem: "How can you quote such large numbers based on such relatively small databases?" Second, the closely related problem of testing robustness because we are combining the evidence by a method whose robustness

we cannot possibly test. Certainly there are strong *a priori* reasons to believe that if all the loci are well separated throughout the genome, then the weight of evidence increases as more and more loci are added. This belief is strengthened by the decreasing proportion of matching profiles found in between-person experiments as more loci are considered. However, the computed statistic, as Stoney (1991) pointed out, is a personal statement of belief, and most certainly not an objective "statistical probability." As an illustration, imagine that we have a 12-locus match for which we have computed a likelihood ratio of 10 billion. We now test an extra four loci, all of which match. Is it now meaningful to say that the likelihood ratio is 100 trillion? Whether or not it is meaningful to quote such an extravagant number, we must be in no doubt that its magnitude depends on independence assumptions to a measure that we cannot possibly support by data. So when we add more loci, the notion that the evidence is becoming more and more compelling is intimately related to personal belief. There is nothing wrong with this, and indeed there is nothing new about it because it is the notion that fingerprint experts invoke as they find more and more points of comparison. The same applies to handwriting comparison, toolmark examination, and so on. So, whereas we sympathize with the view that the weight of the evidence increases with an increasing number of matching loci, we must accept that an assessment made on the basis of a large number of loci is necessarily subjective. We also note that further work is needed to quantify the effects of multilocus associations (Donnelly 1995; J.S. Buckleton, personal communication).

Statistical Basis for Individualization?

One of the questions often asked is just how small the DNA match probability should be, or how large should the likelihood ratio be, to prove a case beyond reasonable doubt "in the absence of other evidence." Several points need to be made. First, it is difficult to envisage a trial in which there really is *no* other evidence. At the very least, there is likely to be evidence relating to the location where the crime occurred and some general area where the defendant might have been expected to be, however ill-defined. A useful discussion of how geographic considerations may be used to formulate prior odds has been given by Walsh et al. (1994). Second, "beyond reasonable doubt" has no numerical standard, so the search for a threshold value for the posterior odds is doomed to fail. Third, the issue is related to that of individuality as it is expressed in other areas of forensic science.

When DNA evidence is presented statistically, it is natural to attempt

to rationalize individualization from statistical lines of reasoning. The 1996 NRC report took a classical statistical view by asking what would be needed to make it very unlikely that a profile occurs in two people. Their argument proceeds as follows. Assume that we find that the suspect has the same genotype as a crime sample. If the match probability is P, then the probability that an unknown person would not be the same genotype is $(1 - P)$. In a population of unrelated people of size N the probability that no one has this genotype is $(1 - P)^N$. The probability that there is at least one other person in the population is $(1 - (1 - P)^N)$. Denote this probability by α:

$$\alpha = 1 - (1 - P)^N$$

For a given value of the population size N, we can calculate the value of P that satisfies a chosen value of α by rearranging the formula:

$$P = 1 - (1 - \alpha)^{1/N} \approx \alpha/N$$

We could arrive at the same numerical value for P by means of the Poisson approximation to the binomial.

For example, if $N = 260$ million, and if we specify that α should be less than 0.001, then P should be less than one in 260 billion. This might be viewed as implying a 99.9% probability that nobody else in the population has the profile. An immediate problem with this formulation is the assumption of independence between all the genotypes in the population, which cannot be true. Furthermore, the apparent objectivity of the calculation is an illusion. Recall our discussion of combining genotypes across loci: We cannot carry out experiments to investigate the robustness of the independence assumptions inherent in a match probability as small as 1 in 260 billion. The number is heavily dependent on subjective judgment; whether it has any real provenance is a matter of opinion.

There is a directly equivalent Bayesian argument based on the idea that every one of the N people in the population has the same chance of being the offender. Then the prior odds in relation to the prosecution proposition are (approximately) $1/N$. The likelihood ratio is $1/P$, so the posterior odds are $1/NP$. Using the above values for N and P we see that the posterior odds are 1000–equivalent to a posterior probability in favor of the prosecution proposition of 0.999. Of course, the scientist is in no position to determine whether 0.999 satisfies the criterion of "beyond reasonable doubt." The NRC report also makes the point that the magnitude of α would be a matter for the court.

We have seen that it is not reasonable to attempt to draw a simple parallel between DNA evidence and a fingerprint identification. The goal of individualization is "beyond reasonable doubt" based on *all* the evidence, and this should be made clear to the court. We see no great problem, providing the evidence justifies it, in an expert saying something of the form, "In my opinion, it is unlikely that there is another person in the country with that profile." How the jury views that opinion and combines it with the non-scientific evidence is another matter. In the UK, at least, it has been heavily implied by the courts that is not for the scientist to consider. However, we have difficulty with a statement of the form, "In my opinion, the crime sample was left by the suspect." This is a quite different kind of opinion and we have seen that such a statement cannot be made without invoking a prior of some kind–unless, like the fingerprint expert, we believe the LR to be so large that it does not matter how small the prior is. In such an instance, perhaps the expert should say, "No matter what evidence the court hears to the contrary, it should accept my opinion that the crime stain came from the suspect." It remains to be seen how the court would react.

DNA EVIDENCE IN COURT: A FUTURE VIEW

We have explained how we believe the statement should be written in the case of Mr. S. We are conscious that we have said more about what the scientist should *not* say if the case comes to court than what he should say. We hope we have said enough to guide him away from the more obvious pitfalls. In this particular case, involving only one sample of transferred material, there is no strong case for preferring a LR over a match probability. How that is combined with the other evidence in the case depends on the court, and the scientist might not be asked for an opinion on this. If he is, however, we believe that there is no more cogent way of explaining the weight of the evidence than by the simple prior/posterior argument, which, after careful explanation to the jury might culminate with an illustration as follows:

> If, based on all the non-DNA evidence the jury consider the odds in favor of the defendant being the assailant to be ten thousand to one against, then the DNA evidence turns those into odds of one hundred to one on.

No doubt it could be argued that the prior of ten thousand to one against might put the wrong idea into the mind of the jury, so it would be better to

elaborate with a range of examples based on a range of illustrative values for the prior. This was suggested in the paternity field by Berry and Geisser (1986), who also provided a useful discussion of how the quantification of priors might be worked at with a jury. There is no doubt that the vast majority of forensic scientists who are currently practicing in this field would be apprehensive about taking this line. This is hardly surprising because it is largely novel and runs counter to whatever training they have received. Courts themselves are steeped in traditional approaches and are naturally conservative. The reactions of British courts demonstrate how far the logical thinking about legal inference has outstripped practice. But this is the way forward. Forensic science ought to be logical, and we have offered a logical approach. Inferentially, practical forensic science has evolved little, if at all, from its dawn–in spite of penetrating insights of Kirk (1963), Kingston (1965), and Stoney (1991). The wonderful new techniques of DNA profiling provide us with an opportunity to progress into the next millenium with the knowledge that forensic science is a true science and with a deeper understanding of its fundamental principles.

SUMMARY

In this final chapter we have shown how the ideas in the preceding chapters are brought together to present DNA profiling evidence in a balanced and robust manner. We have shown that the Bayesian view of evidence provides principles for a scientific approach to interpretation and how those principles provide a basis for balanced and comprehensible reporting. We have discussed the challenges associated with presenting evidence at court and have explained the nature of the fallacious lines of reasoning that the scientist may encounter.

We have considered the issues that are central to the concept of forensic individualization and have talked about the differences in approach between fingerprint and DNA experts. For non-DNA evidence, individualization depends on a leap of faith that, in Bayesian terms, is equivalent to saying "My personal likelihood ratio is so large that, no matter how small the prior odds are, the posterior odds are large enough for me to individualize with certainty." For DNA evidence such a state could be reached only by testing at more and more loci, but the apparent objectivity of numerical statements then becomes increasingly illusory, and the element of personal belief totally dominates the data.

Appendix A

Statistical Tables

Table A.1: Areas under the standard normal density curve, beyond value z.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
4.0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table A.2: Chi-square values that are exceeded with specified probabilities.

	Probability									
d.f.	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21	10.6
3	0.07	0.12	0.22	0.35	0.58	6.25	7.81	9.35	11.3	12.8
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.1	13.3	14.9
5	0.41	0.55	0.83	1.15	1.61	9.24	11.1	12.8	15.1	16.7
6	0.68	0.87	1.24	1.64	2.20	10.6	12.6	14.4	16.8	18.5
7	0.99	1.24	1.69	2.17	2.83	12.0	14.1	16.0	18.5	20.3
8	1.34	1.65	2.18	2.73	3.49	13.4	15.5	17.5	20.1	22.0
9	1.73	2.09	2.70	3.33	4.17	14.7	16.9	19.0	21.7	23.6
10	2.16	2.56	3.25	3.94	4.87	16.0	18.3	20.5	23.2	25.2
11	2.60	3.05	3.82	4.57	5.58	17.3	19.7	21.9	24.7	26.8
12	3.07	3.57	4.40	5.23	6.30	18.5	21.0	23.3	26.2	28.3
13	3.57	4.11	5.01	5.89	7.04	19.8	22.4	24.7	27.7	29.8
14	4.07	4.66	5.63	6.57	7.79	21.1	23.7	26.1	29.1	31.3
15	4.60	5.23	6.26	7.26	8.55	22.3	25.0	27.5	30.6	32.8
16	5.14	5.81	6.91	7.96	9.31	23.5	26.3	28.8	32.0	34.3
17	5.70	6.41	7.56	8.67	10.1	24.8	27.6	30.2	33.4	35.7
18	6.26	7.01	8.23	9.39	10.9	26.0	28.9	31.5	34.8	37.2
19	6.84	7.63	8.91	10.1	11.7	27.2	30.1	32.9	36.2	38.6
20	7.43	8.26	9.59	10.9	12.4	28.4	31.4	34.2	37.6	40.0
21	8.03	8.90	10.3	11.6	13.2	29.6	32.7	35.5	38.9	41.4
22	8.64	9.54	11.0	12.3	14.0	30.8	33.9	36.8	40.3	42.8
23	9.26	10.2	11.7	13.1	14.8	32.0	35.2	38.1	41.6	44.2
24	9.89	10.9	12.4	13.8	15.7	33.2	36.4	39.4	43.0	45.6
25	10.5	11.5	13.1	14.6	16.5	34.4	37.7	40.6	44.3	46.9
26	11.2	12.2	13.8	15.4	17.3	35.6	38.9	41.9	45.6	48.3
27	11.8	12.9	14.6	16.2	18.1	36.7	40.1	43.2	47.0	49.6
28	12.5	13.6	15.3	16.9	18.9	37.9	41.3	44.5	48.3	51.0
29	13.1	14.3	16.0	17.7	19.8	39.1	42.6	45.7	49.6	52.3
30	13.8	15.0	16.8	18.5	20.6	40.3	43.8	47.0	50.9	53.7
40	20.7	22.2	24.4	26.5	29.1	51.8	55.8	59.3	63.7	66.8
50	28.0	29.7	32.4	34.8	37.7	63.2	67.5	71.4	76.2	79.5
60	35.5	37.5	40.5	43.2	46.5	74.4	79.1	83.3	88.4	92.0
70	43.3	45.4	48.8	51.7	55.3	85.5	90.5	95.0	100.4	104.2
80	51.2	53.5	57.2	60.4	64.3	96.6	101.9	106.6	112.3	116.3
100	67.3	70.1	74.2	77.9	82.4	118.5	124.3	129.6	135.8	140.2

Table A.3: Two thousand random digits.

	5	10	15	20	25	30	35	40	45	50
1	30246	86149	45548	80480	85924	02411	46456	23952	55145	18300
2	02806	20733	30853	08034	21238	39933	90958	87912	82486	96960
3	84868	17425	91536	08208	44761	40101	74109	08696	73249	10885
4	65043	86343	36953	04658	42008	84984	49584	53872	52737	24217
5	59792	12608	73246	57277	29384	02608	78779	59311	08421	72618
6	29008	02705	38780	09675	32573	74039	85654	12731	36846	21341
7	74800	20695	99211	38699	28454	21400	11524	81212	55327	93367
8	45715	29459	60745	64762	81553	00401	21852	65586	51269	73813
9	70056	78054	16563	32244	81117	26808	94318	00873	00154	81690
10	30072	38515	52181	21872	17193	57361	16000	51633	70345	48725
11	19490	00789	48629	84877	18858	73868	05461	57469	58009	23998
12	79558	05067	71799	72777	45475	39847	14211	09764	38988	94242
13	18072	34286	46778	95843	31600	57151	89995	58712	46820	81464
14	09933	43223	27657	00697	84736	96171	18120	74205	86558	72670
15	68396	26040	44227	73036	11903	59352	73105	88131	25523	48473
16	76023	01624	74545	18347	66573	79479	24729	98822	93629	72477
17	52257	64895	96218	45817	93951	30547	93632	21510	17326	95743
18	27531	76301	89645	24680	93157	56419	92677	05539	81408	37221
19	17406	68465	66526	13785	92655	25101	95658	54255	07336	17904
20	87810	83955	12467	83985	39484	80179	96878	67468	16173	29937
21	01109	37024	09219	04303	65058	07201	50126	56572	97194	99595
22	67362	79269	61078	70412	89414	45697	17368	48025	41999	45286
23	38002	58000	50220	34603	73647	06894	84712	52922	73303	22802
24	60044	14258	82451	24551	14223	77858	61729	69565	62211	90630
25	55818	55177	80015	88181	96369	57150	37206	02369	18457	29621
26	82646	47169	71375	65259	13194	59086	81076	08421	47402	25764
27	47133	75669	28424	83710	21907	46183	21782	04475	88099	33155
28	62065	06444	34797	56543	90176	41665	53588	71810	26557	83977
29	52765	89407	17693	33927	97348	72061	14231	12340	44493	64194
30	68651	84960	60535	51369	08459	97693	31991	37836	37247	50762
31	74437	48122	89309	16025	06062	10840	22809	28746	30682	48082
32	49051	14405	76357	57632	46511	00666	09647	61493	66875	29164
33	95023	70370	60841	58975	63641	71478	48327	82378	17689	49232
34	19358	28765	57897	93980	61832	10202	79416	40162	85205	87337
35	95489	73778	86660	39424	89005	68527	85534	77132	95116	65790
36	07758	15002	18281	35417	07440	56681	31392	91160	85337	79306
37	27602	69590	13299	50384	25829	85184	89773	97149	16399	41287
38	75864	68804	37205	39021	67019	38964	62848	40359	22254	54700
39	47313	78390	64495	14918	97584	73636	55745	33592	16050	86578
40	13406	80860	65073	73149	74121	97974	60190	50744	52846	91673

Appendix B

SOLUTIONS TO EXERCISES

CHAPTER 1

Exercise 1.1

$$\begin{aligned}
\Pr(H_1|E) &= \frac{1}{3} \\
\Pr(H_1 \textit{ or } H_2|E) &= \frac{1}{3} + \frac{1}{3} = \frac{2}{3} \\
\Pr(\bar{H}_1|E) &= 1 - \frac{1}{3} = \frac{2}{3} \\
\Pr(H_1|0) &= 0
\end{aligned}$$

Exercise 1.2

Let C denote Caucasian, H denote highland, L denote Celtic language, and E denote the information that the person is selected at random from the voter registration list. Then

$$\begin{aligned}
\Pr(C, H, L|E) &= \Pr(L|C, H, E) \Pr(H|C, E) \Pr(C|E) \\
&= 0.75 \times 0.2 \times 0.8 \\
&= 0.12
\end{aligned}$$

Exercise 1.3

Let G denote the event that a person has the required genotype, Ca the event that a person is Caucasian, Mo the event that a person is Maori, and Pa the event that

a person is Pacific Islander. Then

$$\begin{aligned}\Pr(G) &= \Pr(G|Ca)\Pr(Ca) + \Pr(G|Mo)\Pr(Mo) + \Pr(G|Pa)\Pr(Pa) \\ &= 0.013 \times 0.8347 + 0.045 \times 0.1219 + 0.039 \times 0.0434 \\ &= 0.018\end{aligned}$$

Exercise 1.4

a.

$$\begin{aligned}\Pr(\text{B}oth\ dice\ are\ even) &= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \\ O(\text{B}oth\ dice\ are\ even) &= \frac{\frac{1}{4}}{1 - \frac{1}{4}} = \frac{1}{3}\end{aligned}$$

i.e., 3 to 1 against.

b.

$$\begin{aligned}\Pr(\text{B}oth\ dice\ show\ a\ six) &= \frac{1}{6} \times \frac{1}{6} = \frac{1}{36} \\ O(\text{B}oth\ dice\ show\ a\ six) &= \frac{\frac{1}{36}}{1 - \frac{1}{36}} = \frac{1}{35}\end{aligned}$$

i.e., 35 to 1 against.

Exercise 1.5

a.

$$\Pr(H) = \frac{19}{1+19} = \frac{19}{20} = 0.950$$

b.

$$\Pr(H) = \frac{0.2}{1+0.2} = \frac{1}{6} = 0.167$$

c.

$$\Pr(H) = \frac{1000}{1+1000} = \frac{1000}{1001} = 0.999$$

d.

$$\Pr(H) = \frac{1/1000}{1+1/1000} = \frac{1}{1001} = 0.001$$

Exercise 1.6

a.

$$\Pr(D|I) \quad = \quad 1/10,000 = 0.0001$$

b.

$$\Pr(E|D,I) \quad = \quad 0.99$$

c.

$$\Pr(E|\bar{D},I) \quad = \quad 0.05$$

d.

$$\begin{aligned} \Pr\textit{ior Odds} &= \frac{\Pr(D|I)}{\Pr(\bar{D})|I} \\ &= \frac{1/10,000}{1-1/10,000} \\ &= 1/9,999 \end{aligned}$$

i.e., approximately 10,000 to 1 against.

e.

$$\begin{aligned} LR &= \frac{\Pr(E|D,I)}{\Pr(E|\bar{D},I)} \\ &= \frac{0.99}{0.05} = 19.8 \end{aligned}$$

f.

$$\Pr\textit{osterior Odds} \quad = \quad 19.8 \times \frac{1}{9,999} = 0.00198$$

The posterior odds are approximately 500 to 1 against X being infected with the disease. Even though X has tested positive, the probability he has the disease is small, approximately 0.002. Whereas $\Pr(E|D,I) = 0.99$, we note that $\Pr(D|E,I) = 0.002$. We meet an analogous situation when we discuss the transposed conditional in Chapter 2.

CHAPTER 2

Exercise 2.1

a. Correct. Use of the word "if" makes the conditioning clear.

b. Correct. However, the meaning is not conveyed as clearly as in **a.**.

c. Wrong. A clear example of the prosecutor's fallacy.

d. Correct. The evidence is described in terms of a likelihood ratio.

e. Wrong. The likelihood ratio has been transposed into a statement of posterior odds.

f. Wrong. Another posterior odds statement.

g. Wrong.

h. Ambiguous. Depends on whether the sentence is read as: "The chance that a man other than Smith would leave blood of this type ...;" or "The chance that a man other than Smith did leave blood of this type ...." The first is correct, and the second is wrong.

i. Correct.

j. Wrong.

k. Wrong. Although not numerical, it is clearly a probability statement about a proposition.

l. Correct. This form of reporting is discussed in Chapter 9.

CHAPTER 3

Exercise 3.1

a. 3 objects chosen from 5:

$$\binom{5}{3} = \frac{5!}{3!2!} = 10$$

b. 5 objects from 12:

$$\binom{12}{5} = \frac{12!}{5!7!} = 792$$

c. 4 objects chosen from 40:

$$\binom{40}{4} = \frac{40!}{4!36!} = 91,390$$

d. 36 objects from 40:

$$\binom{40}{36} = \frac{40!}{36!4!} = 91,390$$

Exercise 3.2

a. 3 black balls chosen from 5 balls:

$$\frac{5!}{3!2!}(0.5)^5 = 0.3125$$

b. 5 objects from 12:

$$\frac{12!}{5!7!}(0.5)^{12} = 0.1934$$

Exercise 3.3

a. $n = 5, p = 1/16$. The probability of two "6"s is

$$\frac{5!}{2!3!}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^3 = 0.1608$$

b. For the probability of more than two "6"s, we note that

$$\begin{aligned}\Pr(zero\ \text{“6”}s) &= \left(\frac{5}{6}\right)^5 = 0.4019\\ \Pr(one\ \text{“6”}s) &= 5\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^4 = 0.4019\end{aligned}$$

Therefore, the probability of two "6"s or less is $0.1608 + 0.4019 + 0.4019 = 0.9646$, and the probability of more than two "6"s must be $1 - 0.9646 = 0.0354$.

Exercise 3.4

The probability that a person's birthday falls in a given week is (approximately) 1/52. For 50 people, in a given week,

$$\Pr(none\ have\ a\ birthday) = \left(\frac{51}{52}\right)^{50} = 0.3787$$

$$\begin{aligned}
\Pr(\textit{one has a birthday}) &= \binom{50}{1}\left(\frac{1}{52}\right)^{1}\left(\frac{51}{52}\right)^{49} = 0.3713 \\
\Pr(\textit{two have a birthday}) &= \binom{50}{2}\left(\frac{1}{52}\right)^{2}\left(\frac{51}{52}\right)^{48} = 0.1784 \\
\Pr(\textit{two or fewer birthdays}) &= 0.3787 + 0.3713 + 0.1784 \\
&= 0.9284 \\
\Pr(\textit{more than two birthdays}) &= 1 - 0.9284 \\
&= 0.0716
\end{aligned}$$

Exercise 3.5

a.

Number of black balls	Number of ways	Probability
0	1	$1 \times (0.9)^6 = 0.531441$
1	6	$6 \times (0.9)^5(0.1)^1 = 0.354294$
2	15	$15 \times (0.9)^4(0.1)^2 = 0.098415$
3	20	$20 \times (0.9)^3(0.1)^3 = 0.014580$
4	15	$15 \times (0.9)^2(0.1)^4 = 0.001250$
5	6	$6 \times (0.9)^1(0.1)^5 = 0.000054$
6	1	$1 \times (0.1)^6 = 0.000001$

b.

Number of black balls	Number of ways	Probability
0	1	$1 \times (0.5)^6 = 0.015625$
1	6	$6 \times (0.5)^6 = 0.093750$
2	15	$15 \times (0.5)^6 = 0.234375$
3	20	$20 \times (0.5)^6 = 0.312500$
4	15	$15 \times (0.5)^6 = 0.234375$
5	6	$6 \times (0.5)^6 = 0.093750$
6	1	$1 \times (0.5)^6 = 0.016525$

Exercise 3.6

With $n = 400$ and $p = 0.8$:

$$\begin{aligned}
\textit{Mean} &= np = 320 \\
\textit{Variance} &= np(1-p) = 64 \\
\textit{sd} &= \sqrt{np(1-p)} = 8
\end{aligned}$$

Exercise 3.7

With $n = 10^6, p = 10^{-6}$, we have $\lambda = np = 1$, so that

No. of men	Probability
0	0.368
1	0.368
2	0.184
More than 2	0.080

Exercise 3.8

From Table 3.5:

Number of black balls	Number of white balls	Number of red balls	Number of ways	Probability
3	0	0	1	$1 \times (0.5)^3 = 0.125$
2	1	0	3	$3 \times (0.5)^2 \times (0.3) = 0.225$
2	0	1	3	$3 \times (0.5)^2 \times (0.2) = 0.150$
1	2	0	3	$3 \times (0.5) \times (0.3)^2 = 0.135$
1	1	1	6	$6 \times (0.5) \times (0.3) \times (0.2) = 0.180$
1	0	2	3	$3 \times (0.5) \times (0.2)^2 = 0.060$
0	3	0	1	$1 \times (0.3)^3 = 0.027$
0	2	1	3	$3 \times (0.3)^2 \times (0.2) = 0.054$
0	1	2	3	$3 \times (0.3) \times (0.2)^2 = 0.036$
0	0	3	1	$1 \times (0.2)^3 = 0.008$

Exercise 3.9

If x_1, x_2, or x_3 are numbers of black, white, or red balls:

x_1	x_2	x_3	$x_1!$	$x_2!$	$x_3!$	$p_i^{x_1}$	$p_2^{x_2}$	$p_3^{x_3}$	$Prob.$
3	0	0	6	1	1	0.125	1.000	1.000	0.125
2	1	0	2	1	1	0.250	0.300	1.000	0.225
2	0	1	2	1	1	0.250	1.000	0.200	0.150
1	2	0	1	2	1	0.500	0.090	1.000	0.135
1	1	1	1	1	1	0.500	0.300	0.200	0.180
1	0	2	1	1	2	0.500	1.000	0.040	0.060
0	3	0	1	6	1	1.000	0.027	1.000	0.027
0	2	1	1	2	1	1.000	0.090	0.200	0.054
0	1	2	1	1	1	1.000	0.300	0.040	0.036
0	0	3	1	1	6	1.000	1.000	0.008	0.008

Exercise 3.10

If p is the probability of a black ball, then $p = 0.5$ and $q = 1 - p = 0.5$.

No. of black balls	Binomial probability	Lines in table for Ex. 3.8	Combined probabilities from Exercise 3.8
0	0.125	7,8,9,10	$0.027 + 0.054 + 0.036 + 0.008 = 0.125$
1	0.375	4,5,6	$0.135 + 0.180 + 0.00 = 0.375$
2	0.375	2,3	$0.225 + 0.150 = 0.375$
3	0.125	1	0.125

Exercise 3.11

a. The sample proportion is $8/50 = 0.16$ and a normal-approximation confidence interval is $0.16 \pm 1.96\sqrt{0.16 \times 0.84/50} = 0.16 \pm 0.10$. This is the interval (0.06,0.26).

b. The sample proportion is $1/50 = 0.02$ and the normal approximation would not be appropriate. We would use bootstrapping to find a confidence interval (Chapter 5). If the true proportion was 0.02, we can use the binomial theorem to show that the probabilities of 0,1,2 or 3 occurrences of the genotype in a sample of size 50 are 0.36, 0.37, 0.19, or 0.06. We might, therefore, take the range (0, 0.04) to be a 92% confidence interval, and the range (0,0.06) to be a 98% confidence interval.

Exercise 3.12

Category	Observed (o)	Expected e	$(o-e)$	$(o-e)^2/e$
Red	7	5	2	0.8
Black	3	5	-2	0.8
Total	10	10	0	1.6

The null hypothesis is not rejected at the 5% level.

Exercise 3.13

x	Probability	Cumulative probability
10 or 0	$2 \times 0.0010 = 0.0020$	0.0020
9 or 1	$2 \times 0.0098 = 0.0195$	0.0215
8 or 2	$2 \times 0.0439 = 0.0879$	0.1094
7 or 3	$2 \times 0.1172 = 0.2344$	0.3438
6 or 4	$2 \times 0.2051 = 0.4102$	0.7540
5	0.2460	1.0000

The P-value is 0.3438, so the null hypothesis is not rejected at the 5% level.

CHAPTER 4

Exercise 4.1

The three allelic proportions are:

$$\begin{aligned} p_1 &= 0.36 + \frac{1}{2}0.36 + \frac{1}{2}0.12 = 0.60 \\ p_2 &= 0.09 + \frac{1}{2}0.36 + \frac{1}{2}0.06 = 0.30 \\ p_3 &= 0.01 + \frac{1}{2}0.12 + \frac{1}{2}0.06 = 0.10 \end{aligned}$$

Exercise 4.2

The three genotypic proportions are:

$$\begin{aligned} P_{11} &= p_1^2 = 0.49 \\ P_{12} &= 2p_1p_2 = 0.42 \\ P_{22} &= p_2^2 = 0.09 \end{aligned}$$

Exercise 4.3

There are six genotypic proportions:

$$\begin{aligned} P_{AA} &= p_A^2 = 0.04 \\ P_{AO} &= 2p_Ap_O = 0.28 \\ P_{BB} &= p_B^2 = 0.01 \\ P_{BO} &= 2p_Bp_O = 0.14 \\ P_{AB} &= 2p_Ap_B = 0.04 \\ P_{OO} &= p_O^2 = 0.49 \end{aligned}$$

This leads to four phenotypic proportions:

$$\begin{aligned} A &: 0.04 + 0.28 = 0.32 \\ B &: 0.01 + 0.14 = 0.15 \\ AB &: 0.04 \\ O &: 0.49 \end{aligned}$$

Exercise 4.4

The average of the five allelic proportions is $\bar{p} = 0.5$, and this remains constant over time. The proportions in populations I, II, III, IV and V after one or two generations are:

Time	I	II	III	IV	V
0	0.1	0.3	0.5	0.7	0.9
1	0.14	0.32	0.50	0.68	0.86
2	0.176	0.338	0.500	0.662	0.824

Exercise 4.5

The following table displays each of the relevant quantities for subpopulations I, II, III and the total population:

	$i = I$	$i = II$	$i = III$	Total
m_i	0.5	0.3	0.2	1
p_{A_i}	0.4	0.6	0.2	
$m_i p_{A_i}$	0.20	0.18	0.04	0.42
P_{AA_i}	0.16	0.36	0.04	
$m_i P_{AA_i}$	0.080	0.108	0.008	0.196

Note that the Hardy-Weinberg proportion for homozygotes in the whole population is $0.42^2 = 0.1764$, which is less than the actual proportion of 0.1960. The data show a homozygote excess.

Exercise 4.7

For the example shown in Exercise 4.6, the five genotypes in the three successive generations are:

0	A_1A_2	A_3A_4	A_5A_6	A_7A_8	A_9A_{10}
1	A_9A_3	$A_{10}A_9$	A_1A_6	$A_{10}A_2$	A_3A_{10}
2	A_6A_{10}	A_6A_3	A_9A_{10}	A_2A_9	$A_{10}A_3$

and none of these individuals has two IBD alleles. All the F's are zero.

Exercise 4.8

There are three non-zero δ measures for uncle $X(a, b)$ and nephew $Y(c, d)$. If c is the allele Y received from the sib of X, from Table 4.8:

$$\delta_0 = 1/2, \;\; \delta_{ac} = 1/4, \;\; \delta_{bc} = 1/4$$

Putting these into Equation 4.17 provides

$$\Pr(G_X = A_1A_2, G_Y = A_1A_2) \;\; = \;\; 2p_1^2p_2^2 + p_1p_2(p_1 + p_2)/2$$

Exercise 4.9

$Relationship$	$G = A_iA_i$	$G = A_iA_j$
$Parent - child$	p_i	$(p_i + p_j)/2$
$Grandparent - grandchild$	$p_i(1 + p_i)/2$	$(p_i + p_j + 4p_ip_j)/4$
$Half sibs$	$p_i(1 + p_i)/2$	$(p_i + p_j + 4p_ip_j)/4$
$Uncle - nephew$	$p_i(1 + p_i)/2$	$(p_i + p_j + 4p_ip_j)/4$
$First\ cousins$	$p_i(1 + 3p_i)/4$	$(p_i + p_j + 12p_ip_j)/8$

Exercise 4.10

The calculations can be set out as

$Proportions$	$m_\alpha = 0.8$	$m_\beta = 0.2$	
A	$p_{A_\alpha} = 0.4$	$p_{A_\beta} = 0.2$	
mp_A	0.32	0.04	$p_A = 0.36$
B	$p_{B_\alpha} = 0.2$	$p_{B_\beta} = 0.6$	
mp_B	0.16	0.12	$p_B = 0.28$
AB	$P_{AB_\alpha} = 0.08$	$P_{AB_\beta} = 0.12$	
mP_{AB}	0.064	0.024	$P_{AB} = 0.088$

The overall linkage disequilibrium is therefore

$$D_{AB} = 0.088 - 0.36 \times 0.28 = -0.0128$$

CHAPTER 5

Exercise 5.1

a. *GYPA*

$Genotype$	o	e	$(o - e)$	$(o - e)^2/e$
AA	31	29.91	1.09	0.04
AB	49	51.19	−2.19	0.09
BB	23	21.90	1.10	0.06
$Total$	103	103.00	0.00	0.19
(df)				(1)

The hypothesis is not rejected at the 5% level.

b. *HBGG*

G*enotype*	o	e	$(o-e)$	$(o-e)^2/e$
AA	30	32.10	-2.10	0.14
AB	55	50.24	4.76	0.45
AC	0	0.56	-0.56	0.56
BB	17	19.66	-2.66	0.36
BC	1	0.44	0.56	0.73
CC	0	0.00	0.00	0.00
T*otal*	103	103.00	0.00	2.24
(df)				(3)

The hypothesis is not rejected at the 5% level.

c. *D7S8*

G*enotype*	o	e	$(o-e)$	$(o-e)^2/e$
AA	31	30.45	0.55	0.01
AB	50	51.10	-1.10	0.02
BB	22	21.45	0.55	0.01
T*otal*	103	103.00	0.00	0.05
(df)				(1)

The hypothesis is not rejected at the 5% level.

d. *Gc*

G*enotype*	o	e	$(o-e)$	$(o-e)^2/e$
AA	4	6.56	-2.56	1.00
AB	11	10.35	0.65	0.04
AC	33	28.52	4.48	0.70
BB	8	4.08	3.92	3.77
BC	14	22.49	-8.49	3.21
CC	33	31.00	2.00	0.13
T*otal*	103	103.00	0.00	8.85
(df)				(3)

The hypothesis is rejected at the 5% level.

CHAPTER 6

Exercise 6.1

a. In Table 6.2, we use the line for $G_C = A_iA_j, G_M = A_iA_k$, and $G_{AF} = A_jA_l$ and see that the LR is $1/(2p_j)$. In this case: $A_i = A, A_j = B, A_k = C$, and $A_l = C$. The fact that $A_k = A_l$ does not affect the calculation. From Table 5.4, we take $p_j = 0.199$, so LR= 2.51.

b. In Table 6.4, we use the line for $G_C = A_iA_j, G_M = A_iA_k$, and $G_{AF} = A_jA_l$ and see that LR is the reciprocal of $2[p_j(1-2\theta_{AT}) + \theta_{AT}]$. Setting $\theta_{AT} = 1/16$ for first cousins, and with $p_j = 0.199$ as before, LR=2.11.

Exercise 6.2

a. In Table 6.2, we use the line for $G_C = A_iA_j, G_M = A_iA_j$, and $G_{AF} = A_jA_k$ and see that the LR is $1/2(p_i + p_j)$. In this case: $A_i = A, A_j = B$ and $A_k = C$. From Table 5.4, we take $p_i = 0.252$ and $p_j = 0.199$, so LR= 1.11.

b. In Table 6.4, we use the line for $G_C = A_iA_j, G_M = A_iA_j$, and $G_{AF} = A_jA_k$ and see that LR is the reciprocal of $2[(p_i+p_j)(1-2\theta_{AT})+\theta_{AT}]$. Setting $\theta_{AT} = 1/8$ for half sibs, and with $p_i = 0.252$ and $p_j = 0.199$ as before, LR= 1.08.

Exercise 6.3

The numerator for LR is

$$\Pr(G_C = A_iA_j|G_M = A_iA_j, G_{AF} = A_iA_j, H_p) \quad = \quad \frac{1}{2}$$

Because there is doubt as to which of the child's alleles is maternal and which is paternal, we need to sum over both possibilities, as stated in the text. The denominator for LR is

$$\begin{aligned} \text{D}en. \quad &= \quad \Pr(A_M = A_i|G_M)\Pr(A_P = A_j|G_M, G_{AF}, H_d) \\ &\quad + \Pr(A_M = A_j|G_M)\Pr(A_P = A_i|G_M, G_{AF}, H_d) \\ &= \quad \frac{1}{2}[\Pr(A_P = A_j|G_M, G_{AF}, H_d) + \Pr(A_P = A_i|G_M, G_{AF}, H_d)] \end{aligned}$$

As in the text, we need to consider the possible genotypes for *MM*, and all we know is that this person must have at least one of alleles A_i and A_j. The analog

of Table 6.5 is:

				T_3	
G_{MM}	$\Pr(G_{MM})$	T_1	T_2	$A_P = A_i$	$A_P = A_j$
A_iA_i	p_i^2	$\frac{1}{2}$	$\frac{p_i^2}{p_i+p_j}$	$\frac{3}{4}$	$\frac{1}{4}$
A_iA_j	$2p_ip_j$	$\frac{1}{2}$	$\frac{2p_ip_j}{p_i+p_j}$	$\frac{1}{2}$	$\frac{1}{2}$
A_jA_j	p_j^2	$\frac{1}{2}$	$\frac{p_j^2}{p_i+p_j}$	$\frac{1}{4}$	$\frac{3}{4}$
A_iA_k $k \neq i,j$	$2p_ip_k$	$\frac{1}{4}$	$\frac{p_ip_k}{p_i+p_j}$	$\frac{1}{2}$	$\frac{1}{4}$
A_jA_k $k \neq i,j$	$2p_jp_k$	$\frac{1}{4}$	$\frac{p_jp_k}{p_i+p_j}$	$\frac{1}{4}$	$\frac{1}{2}$

$\Pr(G_M = A_iA_j | G_{AF} = A_iA_j) = (p_i + p_j)/2$
$T_1 = \Pr(G_M = A_iA_j | G_{MM}, G_{AF} = A_iA_j, H_d)$
$T_2 = \Pr(G_{MM} | G_M, G_{AF})$
$T_3 = \Pr(A_P | G_{MM}, G_{AF} = A_iA_j, H_d)$

Averaging over the two values for A_M, the denominator becomes

$$\begin{aligned} \text{Den.} &= \frac{p_i^2}{2(p_i+p_j)} + \frac{p_ip_j}{p_i+p_j} + \frac{p_j^2}{2(p_i+p_j)} \\ &\quad + \sum_{k\neq i,j} \frac{3p_ip_k}{8(p_i+p_j)} \\ &\quad + \sum_{k\neq i,j} \frac{3p_jp_k}{8(p_i+p_j)} \\ &= \frac{3+p_i+p_j}{8} \end{aligned}$$

and LR is $4/(3 + p_i + p_j)$.

Exercise 6.4

Under the proposition H_p that the sample is from the missing person, the probability of the evidence is:

$$\begin{aligned} \Pr(E|H_p) &= \Pr(G_P, \{G_S\}, G_M, G_C, G_X | H_p) \\ &= \Pr(G_C|G_X, G_M) \Pr(G_X, G_M, \{G_S\}|G_P) \Pr(G_P) \\ &= \Pr(G_C|G_X, G_M) \Pr(G_M) \Pr(G_X, \{G_S\}|G_P) \Pr(G_P) \\ &= \Pr(G_C|G_X, G_M) \Pr(G_M) \Pr(G_P) \sum_{G_F} \Pr(G_X, \{G_S\}|G_P, G_F) \Pr(G_F) \end{aligned}$$

where

$$\begin{aligned}
\Pr(G_C = A_3A_5 | G_X = A_3A_3, G_M = A_5A_6) &= 1/4 \\
\Pr(G_M) &= 2p_5p_6 \\
\Pr(G_P) &= 2p_3p_4 \\
\Pr(G_X, \{G_S\} | G_P = A_3A_4, G_F = A_2A_3) &= 1/1024 \\
\Pr(G_X, \{G_S\} | G_P = A_3A_4, G_F = A_2A_4) &= 0 \\
\Pr(G_F = A_2A_3) &= 2p_2p_3 \\
\Pr(G_F = A_2A_4) &= 2p_2p_4
\end{aligned}$$

Therefore

$$\Pr(E|H_p) = p_2p_3^2p_4p_5p_6/512$$

Under the proposition H_d that the sample is not from the missing person, the probability of the evidence is:

$$\begin{aligned}
\Pr(E|H_p) &= \Pr(G_P, \{G_S\}, G_M, G_C, G_X | H_p) \\
&= \Pr(G_X)\Pr(G_M)\Pr(G_C|G_M)\Pr(\{G_S\}|G_P)\Pr(G_P) \\
&= \Pr(G_X)\Pr(G_M)\Pr(G_P)\sum_{G_F}\Pr(G_C|G_M, G_F)\Pr(\{G_S\}|G_P, G_F)\Pr(G_F)
\end{aligned}$$

where

$$\begin{aligned}
\Pr(G_X) &= p_3^2 \\
\Pr(G_M) &= 2p_5p_6 \\
\Pr(G_P) &= 2p_3p_4 \\
\Pr(G_C = A_3A_5 | G_M = A_5A_6, G_F = A_2A_3) &= 1/4 \\
\Pr(G_C = A_3A_5 | G_M = A_5A_6, G_F = A_2A_4) &= 1/8 \\
\Pr(\{G_S\} | G_P = A_3A_4, G_F = A_2A_3) &= 1/1024 \\
\Pr(\{G_S\} | G_P = A_3A_4, G_F = A_2A_4) &= 1/1024 \\
\Pr(G_F = A_2A_3) &= 2p_2p_3 \\
\Pr(G_F = A_2A_3) &= 2p_2p_4
\end{aligned}$$

Therefore

$$\Pr(E|H_d) = p_2p_3^2p_4p_5p_6(2p_3 + p_4)/1024$$

and LR is

$$LR = 2/(2p_3 + p_4)$$

CHAPTER 7

Exercise 7.1

In each case the two propositions are

H_p: The victim and suspect were the contributors.
H_d: The victim and an unknown person were the contributors.

The calculation can be set out in this table:

G_V	G_S	E_C	$\Pr(E_C\|G_V, G_S, H_p)$	$\Pr(E_C\|G_V, H_d)$	LR
A_1A_1	A_1A_1	A_1A_1	1	p_1^2	$1/p_1^2$
A_1A_2	A_1A_2	A_1A_2	1	$p_1^2 + 2p_1p_2 + p_2^2$	$1/(p_1+p_2)^2$
A_1A_1	A_1A_2	A_1A_2	1	$2p_1p_2 + p_2^2$	$1/p_2(2p_1+p_2)$
A_1A_2	A_1A_1	A_1A_2	1	$p_1^2 + 2p_1p_2 + p_2^2$	$1/(p_1+p_2)^2$
A_1A_1	A_2A_2	A_1A_2	1	$2p_1p_2 + p_2^2$	$1/p_2(2p_1+p_2)$

Exercise 7.2

In this case $E_C = A_1, A_2, A_3$ and $G_S = A_2A_2$. The two propositions are

H_p: The suspect and an unknown person were the contributors.
H_d: Two unknown people were the contributors.

Under H_p, the unknown person must be A_1A_3, so the evidence probability is

$$\Pr(E_C|G_S, H_p) = 2p_1p_3$$

Under H_d, the two unknown people can be A_1A_2, A_2A_3, or A_2A_2, A_1A_3, or A_3A_3, A_1A_2, or A_1A_2, A_1A_3, or A_1A_2, A_2A_3, or A_1A_3, A_2A_3. The evidence probability is

$$\begin{aligned}\Pr(E_C|H_d) &= 4p_1^2p_2p_3 + 4p_1p_2^2p_3 + 4p_1p_2p_3^2 \\ &\quad + 8p_1^2p_2p_3 + 8p_1p_2^2p_3 + 8p_1p_2p_3^2 \\ &= 12p_1p_2p_3(p_1+p_2+p_3)\end{aligned}$$

and LR is $1/6p_2(p_1+p_2+p_3)$.

Aitken, C.G.G. 1995. *Statistics and the Evaluation of Evidence for Forensic Scientists.* Wiley, New York.

Allison, A.C. 1954. Protection afforded by sickle-cell trait against subtertian malarial infection. Brit. Med. J. 1:290–294.

Balding, D.J. 1995. Estimating products in forensic identification using DNA profiles. J. Am. Stat. Assoc. 90:839–844.

Balding, D.J. and P. Donnelly. 1995. Inference in forensic identification. J. Royal Stat. Soc. (Series A). 158: 21–53.

Balding, D.J. and P. Donnelly. 1996. Evaluating DNA profile evidence when the suspect is identified through a database search. J. Forensic Sci. 41:603–607.

Balding, D.J. and R.A. Nichols. 1994. DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. Forensic Sci. Int. 64:125–140.

Balding, D.J. and R.A.Nichols. 1995. A method for characterizing differentiation between populations at multiallelic loci and its implications for establishing identity and paternity. Genetica 96:3–12.

Balding, D.J. and R.A. Nichols. 1997. Significant genetic correlations among Caucasians at forensic DNA loci. Heredity 78:583–589.

Bernardo, J. and A.F.M. Smith. 1994. *Bayesian Theory.* Wiley, Chichester.

Berry, D.A. 1996. *Statistics: A Bayesian perspective.* Wadsworth, Belmont CA.

Berry, D.A. and S. Geisser. 1986. Inference in cases of disputed paternity. In: M.H. DeGroot, S.E. Fienberg and J.B. Kadane (eds). *Statistics and the Law.* Wiley, New York.

Brenner, C. 1997. Symbolic kinship program. Genetics 145:535–542.

Briggs, T.J. 1978. The probative value of bloodstains on clothing. Medicine, Science and the Law 18:79–83.

Buckleton, J.S., K.A.J. Walsh, G.A.F. Seber and D.G. Woodfield. 1987. A stratified approach to the compilation of blood group frequency surveys. J. Forensic Sci. Soc. 27:103–122.

Budowle, B. and K.L. Monson. 1993. The forensic significance of various reference population databases for estimating the variable number of tandem repeat (VNTR) loci profiles. Pp 177–192 in *DNA Fingerprinting: State of the Science*, S.D.J. Pena, R. Chakraborty, J.T. Epplen and A.J. Jeffreys (eds). Birkhäuser, Basel.

Cavalli-Sforza, L.L. 1998. The DNA revolution in population genetics. Trends in Genetics 14:60–65.

Cavalli-Sforza, L.L., P. Menozzi and A. Piazza. 1994. *The History and Geography of Human Genes.* Princeton University Press, Princeton.

Chakraborty, R., M.I. Kamboh, M. Nwankwo and R.E. Ferrell. 1992. Caucasian genes in American blacks: new data. Am. J. Hum. Genet. 50:145–155.

Chakraborty, R., M.R. Srinivasan and S.P. Daiger. 1993. Evaluation of standard error and confidence interval of estimated multilocus genotype probabilities, and their implications in DNA forensics. Am. J. Hum. Genet. 52:60–70.

Cockerham, C.C. 1969. Variance of gene frequencies. Evolution 23:72–84.

Cockerham, C.C. 1971. Higher order probability functions of identity of alleles by descent. Genetics 69:235–246.

Cockerham, C.C. 1973. Analyses of gene frequencies. Genetics 74:679–700.

Cockerham, C.C and B.S. Weir. 1987. Correlations, descent measures: Drift with migration and mutation. Proc. Natl. Acad. Sci. USA 84:8512–8514.

Curran, J.M., C.M. Triggs, J.S. Buckleton and B.S. Weir. 1999. Interpreting DNA mixtures in structured populations. (submitted)

Dawid, A.P. 1994. The island problem: coherent use of identification evidence. Pp 159–170 in *Aspects of Uncertainty: a Tribute to D.V. Lindley.*, P.R. Freeman and A.F.M. Smith (Editors). Wiley, Chichester.

Dawid, A.P. and J. Mortera. 1996. Coherent analysis of forensic identification evidence. J. Royal Stat. Soc. (Series B). 58: 425–433.

Donnelly, P. 1995. Nonindependence of matches at different loci in DNA profiles: quantifying the effect of close relatives on the match probability. Heredity 75:26–34.

Edwards, A.W.F. 1992. *Likelihood.* Expanded Edition. Johns Hopkins Press, Baltimore.

Efron, B. 1982. *The Jackknife, the Bootstrap and Other Resampling Plans.* CBMS-NSF Regional Conference Series in Applied Mathematics, Monograph 38. SIAM, Philadelphia.

Efron, B. and R.J. Tibshirani. 1993. *An Introduction to the Bootstrap.* Chapman and Hall, New York.

Eggleston, R. 1983. *Evidence, Proof and Probability.* Weidenfeld and Nicholson. London.

Essen-Möller, E. 1938. Die Beweiskraft der Aehnlichkeit im Vaterschaftsnachweis; theoretische Grundlagen. Mitt. Anthrop. Ges. (Wien) 68:9–53.

Evett, I.W. 1983. What is the probability that this blood came from that person: A meaningful question? J. Forensic Sci. Soc. 23:35–59.

Evett, I.W. 1995. Avoiding the transposed conditional. Science and Justice 158:41–42.

Evett, I.W. and J.S. Buckleton. 1989. Some aspects of the Bayesian approach to evidence evaluation. J. Forensic Sci. Soc. 29:317–324.

Evett, I.W. and J.S. Buckleton. 1996. Statistical analysis of STR data. Pp. 79–86 in Carracedo, A., B. Brinkmann and W. Bar (Eds.) *Advances in Forensic Haemogenetics* Vol. 6. Springer-Verlag, Berlin.

Evett, I.W., C. Buffery, G. Willott and D. Stoney. 1991. A guide to interpreting single locus profiles of DNA mixtures in forensic case. J. Forensic Sci. Soc. 31:41–47.

Evett, I.W., P. Gill and J.A. Lambert. 1998. Taking account of peak areas when interpreting mixed DNA profiles. J. Forensic Sci. 43:62–69.

Evett, I.W., P.D. Gill, J.K. Scranage and B.S. Weir. 1996a. Establishing the robustness of STR statistics for forensic applications. Am. J. Hum. Genet. 58:398–407.

Evett, I.W., J.A. Lambert, J.S. Buckleton and B.S. Weir. 1996b. Statistical analysis of a large file of STR profiles of British Caucasians to support forensic

casework. Int. J. legal Med. 109:173–177.
Evett, I.W. and B.S. Weir. 1991. Flawed reasoning in court. Chance 4:19–21.
Finkelstein, M.O. and W.B. Fairley. 1970. A Bayesian approach to identification evidence. Harvard Law Review 83:489–517.
Fisher, R.A. 1935. The logic of scientific inference. J. Roy. Stat. Soc. 98:39–54.
Friedman, R.D. 1996. Assessing evidence. Michigan Law Review 94:1810–1838.
Foreman, L.A., A.F.M. Smith and I.W. Evett. 1997. Bayesian analysis of deoxyribonucleic acid profiling data in forensic identification applications. J. Roy. Stat. Soc. 160:
Gill, P. and I.W. Evett. 1995. Population genetics of short tandem repeat loci. Genetica 96:69–87.
Goodman, L.A. 1960. On the exact variance of products. J. Am. Stat. Assoc. 57:54–60.
Hill, W.G., H.A. Babiker, L.C. Ranford-Cartwright and D. Walliker. 1995. Estimation of inbreeding coefficients from genotypic data on multiple alleles, and application to estimation of clonality in malaria parasites, Genet. Res. 65:53–61.
Inman, K. and N. Rudin. 1997. *An Introduction to Forensic DNA Analysis.* CRC Press, Boca Raton FL.
Jaynes, E.T. 1989. Clearing up mysteries–the original goal. Pp. 1–27 in Skilling, J. (Ed.), *Maximum Entropy and Bayesian Methods*. Kluwer, Dordrecht.
Kac, M. 1983. What is random? Am. Scientist 71:405–406.
Kaye, D.H. 1990. DNA paternity probabilities. Fam. Law Q. 24:279–304.
Kind, S.S. 1994. Crime investigation and the criminal trial: a three chapter paradigm of evidence. J. Forensic Sci. Soc. 34:155–164.
Kingston, C.R. 1965. Applications of probability theory in criminalistics. J. Am. Stat. Assoc. 60:70–80.
Kirk, P.L. 1963. The ontogeny of criminalistics. J. Law, Criminology and Police Science 54:235–238.
Lambert, J.A., J.K. Scranage and I.W. Evett. 1995. Large scale database experiments to assess the significance of matching DNA profiles. Int. J. Legal Med. 108:8-13.
Lewontin, R.C. 1993. Which population? Am. J. Hum. Genet. 52:205.
Lewontin, R.C. and C.C. Cockerham. 1959. The goodness-of-fit test for detecting natural selection in random mating populations. Evolution 13:561–564.
Li, C.C. 1988. Pseudo-random mating populations. In celebration of the 80th anniversary of the Hardy-Weinberg law. Genetics 119:731–737.
Li, Y.J. 1996. *Characterizing the Structure of Genetic Populations.* Ph.D Thesis. N.C. State University.
Lindley, D.V. 1977. A problem in forensic science. Biometrika 64:207–213.
Lindley, D.V. 1982. Coherence. Pp. 29–31 in Kotz, S., N.L. Johnson and C.B. Read (Eds.) *Encyclopedia of Statistical Sciences*, Vol. 2. Wiley, New York.
Lindley, D.V. 1991. Probability. Pp. 27–50 in Aitken, C.G.G. and D.A. Stoney (Eds.) *The Use of Statistics in Forensic Science.* Ellis Horwood, New York.
Maiste, P.J. and B.S. Weir. 1995. A comparison of tests for independence in the

FBI RFLP data bases. Genetica 96:125–138.

Morris, J.W., R.A Garber, J. d'Autremont and C.H. Brenner. 1988. The avuncular index and the incest index. Pp. 607–611 in *Advances if Forensic Haemogenetics 1.* Springer-Verlag, Berlin.

Mosteller, F. and D.L. Wallace. 1964. *Applied Bayesian and Classical Inference - the Case of the Federalist Papers.* Springer-Verlag, New York.

National Research Council. 1996. *The Evaluation of Forensic DNA Evidence.* National Academy Press, Washington, DC.

Nichols, R.A. and D.J. Balding. 1991. Effects of population structure on DNA fingerprint analysis in forensic science. Heredity 66:297–302.

O'Hagan, A. 1994. *Kendall's Advanced Theory of Statistics 2B*, Wiley, New York.

Risch, N. and B. Devlin. 1992. On the probability of matching DNA fingerprints. Science 255:717–720.

Robertson, B. and G.A. Vignaux. 1995. *Interpreting Evidence: Evaluating Forensic Science in the Courtroom.* Wiley, Chichester.

Roychoudhury, A.K. and M. Nei. 1988. *Human Polymorphic Genes.* World distribution. Oxford University Press, Oxford.

Stoney, D.A. 1991. What made us ever think we could individualize using statistics? J. For. Sci. Soc. 31: 197–199.

Thompson, W.C. and E.L. Schumann. 1987. Interpretation of statistical evidence in criminal trials - The prosecutors fallacy and the defense attorneys fallacy. Law and Human Behavior 11; 167–187

Tribe, L.H. 1971. Trial by Mathematics: Precision and ritual in the legal process. Harvard Law Review 84:1329–1393.

Vogel, F. and A.G. Motulski. 1986. *Human Genetics,* Second Edition. Springer-Verlag, New York.

Walker, R.H., R.J. Duquesnoy, E.R. Jennings, H.D. Krause, C.L. Lee, and H. Polesky (Eds.). 1983. *Inclusion Probabilities in Parentage Testing.* American Association of Blood Banks. Arlington, VA.

Walsh, K.A.J, J.S. Buckleton and C.M. Triggs. 1994. Assessing prior probabilities considering geography. J. Forensic Sci. Soc. 34: 47-51.

Wambaugh, J. 1989. *The Blooding.* William Morrow, New York.

Weir, B.S. 1994. Effects of inbreeding on forensic calculations. Ann. Rev. Genet. 28:597–621.

Weir, B.S. 1996. *Genetic Data Analysis II.* Sinauer, Sunderland, MA.

Weir, B.S. and C.C. Cockerham. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370.

Weir, B.S. and I.W. Evett. 1992. Whose DNA? Am. J. Hum. Genet. 50:869.

Weir, B.S. and I.W. Evett. 1993. Reply to Lewontin. Am. J. Hum. Genet. 52:206.

Weir, B.S., C.M. Triggs, L. Starling, L.I. Stowell, K.A.J. Walsh and J.S. Buckleton. 1997. Interpreting DNA mixtures. J. Forensic Sci. 42:113–122.

Wright, S. 1951. The genetical structure of populations. Ann. Eugen. 15:32–354.

Wright, S. 1965. The interpretation of population structure by F-statistics with special regard to systems of mating. Evolution 19:395-420.

Zaykin, D., L. Zhivotovsky and B.S. Weir. 1995. Exact tests for association between

alleles at arbitrary numbers of loci. Genetica 96:169–178.