Probability and Statistics for Bioinformatics and Genetics

Course Notes

Paul Maiste
The Johns Hopkins University

These notes are meant solely to be used as study notes for students in
550.435 - Bioinformatics and Statistical Genetics, Spring 2006.

May 2, 2006

Contents

Chapter 1

Introduction to Cell Biology and Genetics

This chapter is an introduction to some of the necessary background in cell biology and genetics. Further topics will be covered throughout the course. Those who have a sound background in these areas already may find it only necessary to skim through this chapter before moving on.

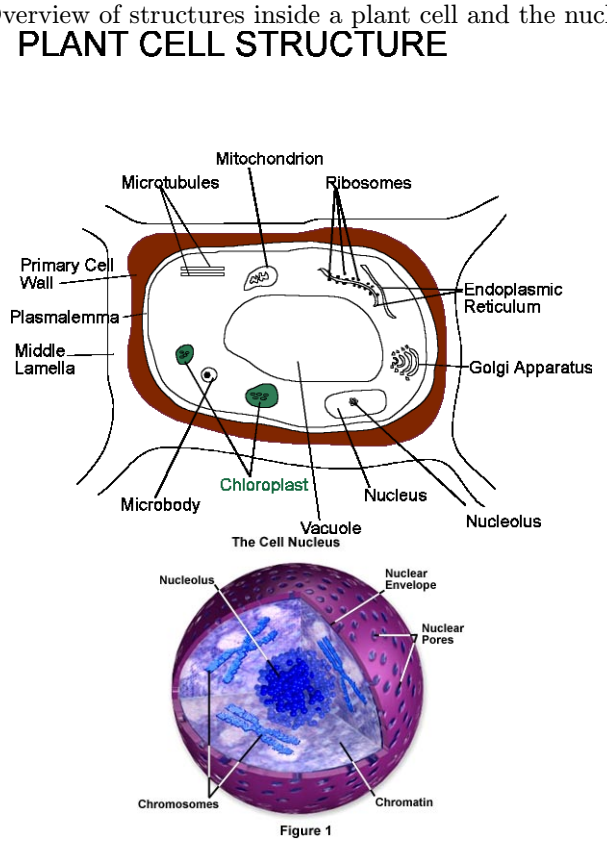
Additionally, we will only cover just the very basic concepts to get started. To get a more comprehensive and detailed background in cell biology and genetics, please refer to the many good texts that have been written on these topics.

1.1 The Cell, DNA, and Chromosomes

The *cell* is the basic structural and functional unit of all organisms. A single cell is the lowest form of life thought to be possible. Most organisms consist of more than one cell, each of which becomes specialised into particular functions towards the cause of the entire organism (such as liver cells, skin cells, etc). Cells possess many structures inside them that contain and maintain the building blocks of life called organelles. Animal cells and plant cells differ fundamentally. DNA exists in the nucleus of the cell. Figure 1.1 shows an overview of the various structures inside a plant cell and the nucleus of a cell.

Deoxyribonucleic acid, or *DNA* is the molecule that encodes genetic information in the nucleus of cells. It determines the structure, function and behaviour of the cell. DNA is a double-stranded molecule held together by weak bonds between *base pairs* of *nucleotides*. The four nucleotides in DNA contain the bases: adenine (A), guanine (G), cytosine (C), and thymine (T). In nature, base pairs form only between A and T and between G and C, thus the base sequence of each single strand can be deduced from that of its partner. In other words, if the nucleotide A is at a certain position on one of the strands of the double-stranded DNA molecule, then the nucleotide T must be on the other strand at that position, with a chemical bond between them. For this reason, even though DNA is actually double stranded, we will only ever need to consider it

Figure 1.1: Overview of structures inside a plant cell and the nucleus of a cell.



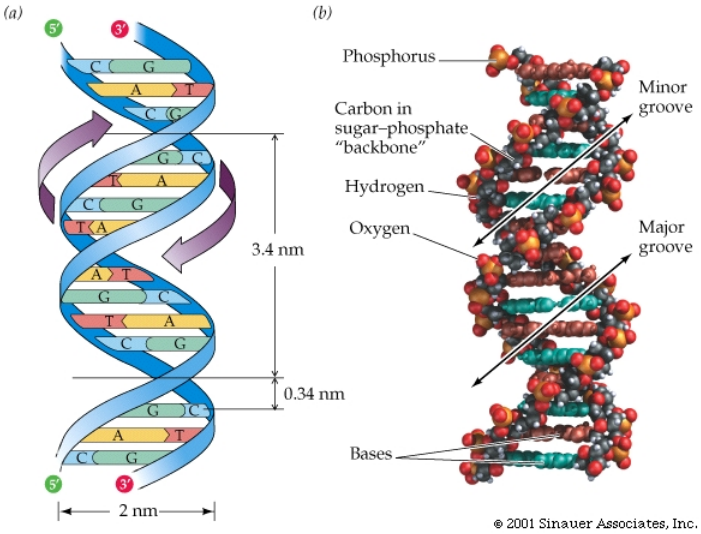
as single stranded in our discussions and analyses.

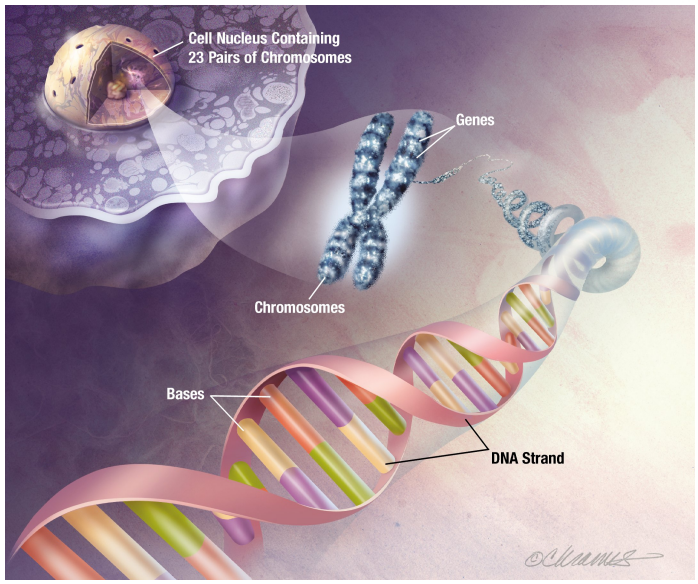
Let's say a little more about some of the terms used above. *Nucleotides* are the basic structural unit of nucleic acids (DNA or RNA which we'll discuss in a little while). There are four nucleotides in nature, represented by the letters A, T, G, and C as mentioned above. The term *base pair* (bp) is often used as a synonym for nucleotide although it specifically refers to the fact that bases (i.e., nucleotides) are paired together across the two strands of DNA. The strands are held together in the shape of a double helix by the bonds between base pairs. Figure 1.2 shows an overview of the double stranded DNA helix.

Other than special cells which we will discuss later, the number of base pairs in a cell is constant for a given organism, but can be very different from one species to another. Table 1.1 shows the differences for some selected organisms. Notice that a typical human cell has approximately three billion base pairs.

The nucleotides that make up DNA in a cell are organized into *chromosomes*. These are self-replicating genetic structures of cells. We can think of them as a linear sequence of nucleotides. Each chromosome has a characteristic length and banding pattern. The number of chromosomes in a cell is different from organism to organism, as seen in Table 1.2.

Figure 1.2: Double Stranded DNA helix





Some cells in an organism are referred to as *haploid cells* and some are referred to as *diploid cells*. Haploid cells have a single copy of each chromosome. Typically, the cells involved in sexual reproduction, called gametes, are haploid. The number of chromosomes in a haploid cell of an organism is referred to as the *haploid number* for that organism. For humans, that number is 23.

Diploid cells, on the other hand, contain two copies, or a pair, of each chromosome. One of the pair of a chromosome comes from the maternal parent of the individual, while the other comes from the paternal parent. Most cells in an animal diploid. The number of chromosomes in a diploid cell of an organism is called the *diploid number* for that organism (and this will always be twice the haploid number). For humans, the diploid number is 46.

Table 1.1: Number of base pairs for selected organisms

Organism	**Estimated total bp**
Bacterium	4,638,858
Yeast	12,069,303
Fruit fly	165,000,000
Plant	100,000,000
Onion	6,000,000,000
Mouse	3,000,000,000
Dog	5,000,000,000
Human	3,000,000,000

Let's focus on diploid cells for a moment. The fact that there is a pair of each chromosome in such a cell is important to understand. When referencing the two chromosomes of a pair, we call them *homologous chromosomes*. The word homologous refers to their very close similarity. They will have the same number of nucleotides, and if they were lined up side by side and the sequence of nucleotides compared, they would be very nearly identical.

Recall that for humans, a diploid cell has 23 pairs of homologous chromosomes, for a total of 46 chromosomes. Of these 23 pairs, 22 are called *autosomes*. Autosomes are not involved in sex determination. For other organisms, the same idea holds: $n - 1$ out of n total pairs are autosomes.

The one other pair of chromosomes is called the *sex chromosomes*. These determine the sex of an animal. In most animals, including humans, females have two copies of the *X chromosome* (referred to as XX), while males have one copy of the X chromosome and one copy of the *Y chromosome* (the combination is referred to as XY).

In the case of females, who are XX, the pair is truly a homologous pair, as with the autosomes. They are of the same length and very nearly the same sequence of nucleotides. In the case of females, the pair is not truly considered homologous, since the X and Y chromosomes are very different in nature. However, we will still refer to them as the "pair" of sex chromosomes.

Birds are an example of animals where the reverse situation determines sex: the male is XX and the female XY. In some organisms, there is only one sex chromosome. One sex is XX and the other is referred to as XO (meaning there is only one chromosome in the pair; the "O" refers to the fact that there is no second chromosome).

Figure 1.3 gives an idea of the natural look of a diploid cell (a human male), showing the 23 pairs of chromosomes and their noticeable banding patterns. The chromosomes in humans (and all species) are numbered from longest to shortest, with the sex chromosomes being unnumbered and referred to using the X and/or

Table 1.2: Haploid and diploid number of various species

Organism	Haploid Chromosome Number	Diploid Chromosome Number
Homo sapiens (human)	23	46
Pan troglodytes (chimpanzee)	24	48
Canis familiaris (dog)	39	78
Felis cattus (cat)	19	38
Equus caballus (horse)	32	64
Equus asinis (ass/donkey)	31	62
Gallus domesticus (chicken)	39	78
Rana pipiens (frog)	13	26
Cyprinus carpio (carp)	52	104
Drosophila melanogaster (fruitfly)	4	8
Pisum sativum (crab)	7	14
Potato	12	24
Ferns	100+	200+

Y notation as above.

Most of the statements made above are not absolute due to the fact that *genetic mutations* can sometimes occur. For example, some individuals may have a different number of chromosomes than the norm for the species, or the chromosomes may be of different length and makeup than the norm. Figure 1.4 gives an overview of some common ways that mutations can change the structure of a chromosome to make it different from the norm. Table 1.3 shows some common human syndromes that are the result of various such structural changes.

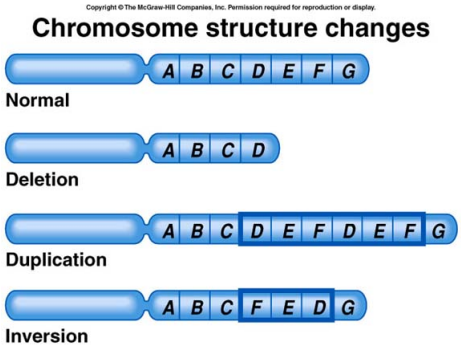
1.2 Function of the Cell and DNA

There are an enormous many important functions that cells and their embedded DNA serve in an individual. Here, we will focus on those functions that we will need as part of the course. Again, please refer to other biology or genetics textbooks for much more detail in this area.

First, recall that for our purposes we will think of a chromosome as a linear sequence of nucleotides. So when we describe a chromosome or a section of a chromosome, it will be with a string of letters, with every letter in the string being one of A,C,G, or T. As mentioned earlier, taken together, there are about 3 billion total letters (i.e., nucleotides) across the 23 human chromosomes that constitute the DNA of an individual. The entire set of DNA in a typical cell in an individual is referred to as its *genome*.

Figure 1.3: 23 pairs of chromosomes in a human male

Figure 1.4: Some common chromosomal abnormalities



Most of this DNA in a cell is what we call *junk DNA*. This is DNA that has no apparent bodily function at all, and in a sense can be considered obsolete. In humans, it is estimated that 90% of the 3 billion base pairs fall into this category of junk DNA. Although such DNA plays no biological role in an individual, we will see that it can still play a very important role in genetic analysis.

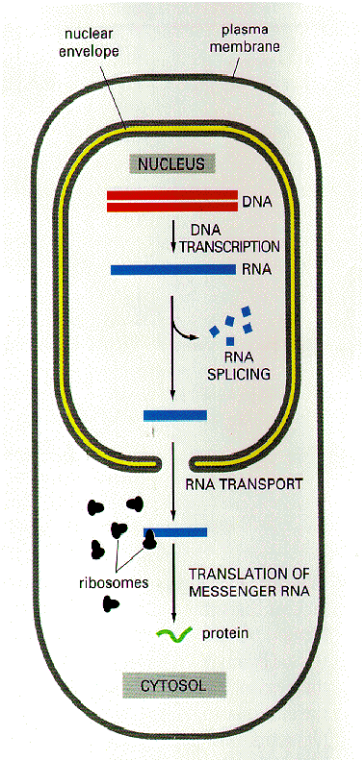
The DNA that does have biological function is organized into what we refer to as *genes*. A gene is a specific location on a chromosome where the DNA contains information required for a particular function, and can be switched on and off on demand. They are responsible for the inherited characteristics that distinguish one individual from another. The region of DNA that makes up a gene is often called a *coding region*. This terminology emphasizes the fact that the underlying DNA nucleotide sequence of a gene is actually a code for creating a certain protein, which we'll see below.

There are many such locations, or genes, in a genome. Table 1.4 shows the estimated number of genes for

selected organisms. Our first instinct would be that humans would contain many times more genes that other organisms since our bodies are vastly more complex than, say, a fly or a plant. However, this table shows that this is not true, and in fact, humans have only 25% more genes than a typical plant, and only twice as many as fruit flies. We do have to remember that the numbers shown in this table are just estimates, and it was not too long ago that geneticists believed that a human genome carried on the order of 100,000 genes.

Now let's spend some time looking more closely at the function of a cell. The ultimate function of the cell is to take genes, as coded in the DNA in the nucleus, and turn them into proteins that are used throughout the cell and the body for various activities. The main functions of the cell can be broken into two areas: (1) Transcription, which happens in the nucleus, and (2) Translation, which happens outside of the nucleus. The two are very intertwined, and we will discuss them below. Figure 1.5 gives an overview of these functions.

Figure 1.5: Overview of cellular function in the nucleus



In short *transcription* can be described as the formation of *RNA* (pre-messenger RNA) from DNA. What is this RNA? For our purposes, we can consider it as very similar to DNA. More technically, RNA is a nucleic acid found in all living cells just like DNA. It plays a role in transferring information from DNA to the protein-forming system of the cell outside of the nucleus. Chemically, it is similar to DNA, though the deoxyribose acid of DNA is replaced with ribose sugar in RNA, and all thymine bases (T) in DNA are replaced with uracil (U) in RNA.

Now let's get back to transcription. To put it simply, a gene is transcribed from DNA to RNA in such a way that the RNA is an exact copy of the DNA, except with T's in DNA replaced by U's in the resulting RNA copy.
How does the cell know where a gene is among all this DNA? There are many complex mechanisms for this. The simplest and most direct thing to point out is that every gene starts with the DNA sequence ATG (resulting in an RNA sequence of AUG).

That doesn't at all mean that every ATG sequence in a genome represents the start of a gene sequence. There are other important factors involved, such as *promotor regions* which guide the necessary molecules to the right location to begin transcription.

Also, a gene always ends with one of three stop sequences in the DNA: TAA, TAG, or TGA. Once one of these three sequences is reached, transcription of DNA to RNA stops.

The actual biochemical process by which this happens is very involved. Below are links to two short animated movie clips on the internet which do a good job of showing the process.

http://www.lakesideschool.org/upperschool/departments/science/bio/17-06-Transcription.mov
http://www.memorial.ecasd.k12.wi.us/Departments/Science/MAllen/Private_Docs/Biotech/Transcription.mov

Ultimately, this RNA copy of a gene makes its way out of the nucleus of the cell to eventually get turned into a certain protein. But before that happens, a couple of other important things happen. First, the RNA molecule that immediately results from a transcription is more correctly referred to as *pre-messenger RNA*. This pre-messenger RNA typically contains *introns*, which are segments of the RNA that get removed, or spliced out after transcription. An intron of a gene is sometimes called a *non-coding region* of the gene, because although it is physically a part of the gene, it is a region that is removed prior to its code can be translated into a protein.

The result of intron removal is *mRNA* or *messenger RNA*. We can think of mRNA as the final result of the transcription process. This molecule is moved out of the nucleus and carries the information that is translated during protein synthesis, which will be discussed below.

A class of organic molecules that plays an important role in cell biology and bioinformatics is *amino acids*. There are twenty amino acids, as shown in Figure 1.6 (along with their common three- and one-letter abbreviations). Amino acids combine in linear arrays to form proteins in living organisms.

Let's say more about proteins. A *protein* is a complex molecule consisting of a particular sequence of amino acids (sometimes called peptides or residues in this context). All proteins consist of carbon, hydrogen, oxygen and nitrogen. Proteins vary in structure according to their function, with the three types of protein being fibrous, globular and conjugated proteins. Each protein has a unique, genetically defined

amino acid sequence which determines its specific shape and function. Proteins serve as enzymes, structural elements, hormones, immunoglobulins, and are involved in oxygen transport, muscle contraction, electron transport and other activities throughout the body and in photosynthesis. It suffices to say that proteins are an extremely important molecule that play a role in the enormous complexities of how an organism functions.

How are proteins created in the cell? After mRNA is carried outside of the cell nucleus, the process of *protein synthesis*, or *translation*, occurs. Protein synthesis is the process in which individual amino acids are connected to each other in a specific order as dictated by the sequence of nucleotides in DNA/RNA. As mentioned above, this sequence of amino acids is the protein.

The biochemical process by which protein synthesis works is quite complex. We will try and get across the important points here:

1. mRNA reaches the cell cytoplasm and is processed in a linear sequence from the start of the mRNA molecule to the end by a complex molecule called a ribosome.
2. The mRNA is processed three bases at a time. A set of three bases is called a *codon* (also referred to as a triplet of bases, or just a triplet). The specific sequence of bases in the codon determines which amino acid is next attached to the growing protein. Which amino acid is attached is determined by the *genetic code* which is shown in Figure 1.7. Since there are $4^3 = 64$ possible codons and only 20 amino acids, it should be noticed that different codons map to the same amino acid (e.g., UUU and UUC both map to the Phe amino acid).
3. Recall from our discussion of transcription above that the first codon (the *start codon*) is always (with rare exceptions) AUG, and the the last codon (the *stop codon*) is always either UAA, UAG, or UGA. The stop codon does not code for an amino acid; it just tells the ribosomal unit where to stop processing the mRNA.
4. After the last amino acid is added to the protein chain and synthesis stops, the protein folds up into a characteristic shape, as determined by the biochemical interactions of the amino acids in the chain. It then goes on its way to serve its function in the cell or the rest of the body.

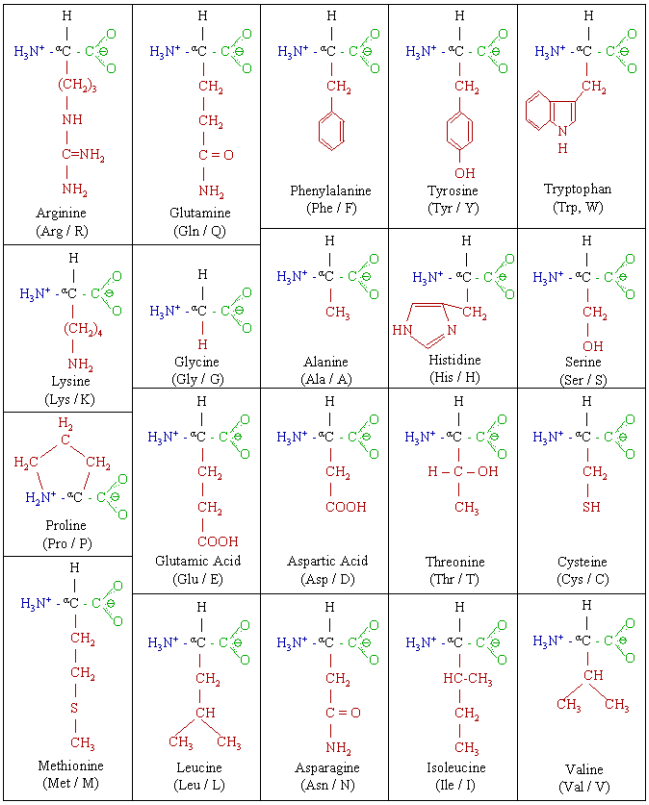
As with transcription, the process of protein synthesis is much more complex than we need to delve into here. Below are links to two animated movie clips that can be found on the web which give a simple yet good overview of the process.

http://www.biology.arizona.edu/molecular_bio/problem_sets/nucleic_acids/graphics/protsyn95.mov
http://www.whfreeman.com/lodish4e/content/video/Protein_Synthesis.mov

1.3 Genetics

With most of our necessary cell biology background complete, we can begin discussing how the field of genetics fits into all of this. We can think of *genetics* as the study of inheritance. Let's review some important

Figure 1.6: The 20 amino acids and their chemical forms



genetic terms.

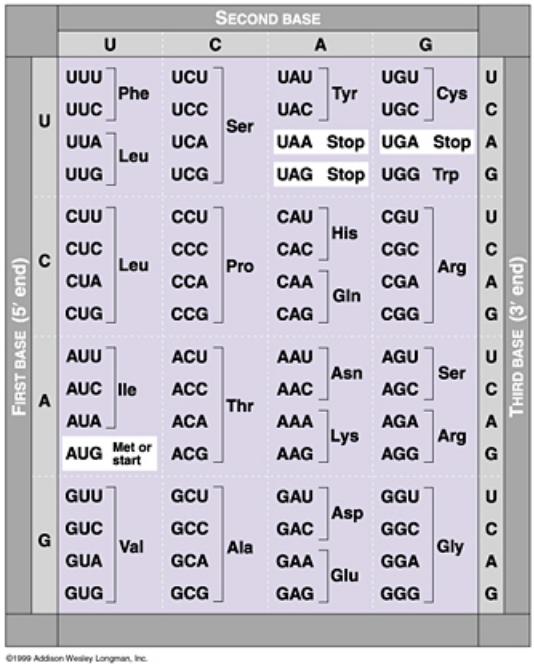
One of the most general terms that we will use throughout the course is *locus* (plural: loci). It is used to refer to a location (any location) in the genome in which we have an interest. For example, a locus that we might be studying is a sequence of 150 bases at a particular spot on the short arm of chromosome 7 in humans. It is a very general term because that location doesn't have to have any specific properties; it could be the location of a gene, or just some other non-coding region of DNA along the chromosome.

Some loci are genes, and some are not. As we've seen earlier, genes are of particular interest to us because they are translated into proteins and so are the key to how a cell and entire organism works. Although a typical genome has many thousands of genes, they are still a rarity within the genome as a whole (we mentioned earlier that in humans, 90% of DNA is not genes).

Even though gene loci are certainly of interest, it is not true that no other regions of DNA are interesting. In fact, regions of DNA outside of genes are extremely useful for genetic analysis, as we'll see, and also in some cases do play a biological role.

A term often used to more specifically refer to a locus that is not a gene is *genetic marker*, or just *marker*.

Figure 1.7: The genetic code that translates RNA triplets into amino acids



Think of a marker as a region of DNA that doesn't necessarily have any biological function, but whose inheritance from one generation to another can be monitored. We'll discuss markers more fully in later chapters.

Now, consider a particular locus that we'll call locus **A** (we'll use boldface capital letters to refer to a generic locus). For the sake of simple discussion, let's say that this locus represents a certain region of DNA consisting of just five bases on chromosome number 3 in humans.

Now consider a particular individual. If we performed an analysis of one copy of that person's chromosome 3, we would be able to view the exact sequence of bases as locus **A**. It might be, say, AATGC. If we then analyzed this locus on this individual's other copy of chromosome 3, the base sequence might be AGTGG.

Further, say we then select another individual and analyze locus **A** on both copies of chromosome 3. We may notice that one of the copies has the sequence AATGC (just like one of the chromosomes of the first individual), but maybe the second copy has the sequence TTTAC (different from any other sequence we've come across so far).

In other words, the DNA sequence that is found at a locus may have alternative forms from one chromosome and one individual to another. These different forms are called *alleles*. In the example above, we found three different alleles for locus **A** among the four chromosomes we analyzed (two for each of the two individuals). If we investigated many more individuals, it is likely that there would be more allelic forms found in the population. Depending on the locus in question, there might be only one allelic form to be found in the

population (every chromosome on every individual has exactly the same DNA sequence at that location), or there may be two allelic forms, or three, or possibly many different alleles.

Let's continue the example above and discuss a little notation. We might refer to the DNA sequence AATGC as Allele 1, or A_1, the sequence AGTGG as Allele 2, or A_2, and the sequence TTTAC as Allele 3, or A_3. In other words, it is common to number the different allelic forms of a locus and give them some sort of shorthand notation like A_1, A_2, and A_3 in this example.

Example 1.1: **ABO Blood group**: Let's look at an example of a real locus, in fact a gene, in humans. The ABO locus can be found on Chromosome 9. It is the well-known gene that determines a persons blood-type. There are three alleles (three different forms of this gene) present in the human population. The three allelic forms are referred to as A, B, and O. These are just shorthand descriptions of the three different sequences of nucleotides that have been found to occur at that location. These alleles produce different protein sequences through transcription of the gene, and consequentially serve slightly different roles with regard to antigens on the surface of red blood cells:

- The A allele produces a protein that serves as a particular antigen (the "A" antigen) that can be found on the surface of red blood cells.
- The B allele produces a different protein that serves as a different antigen (the "B" antigen) that can be found on the surface of red blood cells.
- The O allele does not produce any such antigens when transcribed and translated into a protein.

■

In genetic analysis, we typically want to describe the alleles carried by an individual at a locus. This is called the individual's *genotype*. For most loci, the genotype of an individual consists of a description of which two alleles that individual carries for that locus (since there are two copies of each chromosome).

Going back to the simple five base long locus we have discussed above, we would say that Individual 1's genotype at locus **A** is A_1A_2, and Individual 2's genotype is A_1A_3. When we write the genotype of an individual, we just list the shorthand notation for the two alleles the individual has. Also, it is conventional to always list the alleles in a standard order, such as numerical order in this example. There is no sense of ordering to the two chromosomes we investigated, so there is no particular meaning to the order we list the alleles in (which is why we might as well list them numerically or alphabetically as the case may be).

As a side note, notice that the one situation (for humans and most species) where a genotype would not consist of a listing of *two* alleles is for analysis of the sex chromosome in males. For example, if the locus we were studying was on the X chromosome, then a male only has one copy of that chromosome in each cell, and so would only have one allele for that locus. Notice that a female studied for that same locus would still

have two alleles representing their genotype since they do have two copies of chromosome X.

Example 1.2: **ABO Blood group**: The possible genotypes that an individual can have for this gene are *AA*, *AB*, *AO*, *BB*, *BO*, and *OO*. These are the only six possible. Again, notice that order doesn't matter when listing the alleles that make up a genotype. So *BO* and *OB* are the same genotype; such an individual has one chromosome with the *B* allele and one with the *O* allele.

■

With regard to genotypes, an individual whose two alleles at a locus are the same is called a *homozygote* at that locus. An individual with two different alleles at a locus is called a *heterozygote* at the locus. For example, an individual whose genotype is *AA* at the ABO locus would be called a homozygote (or said to be "homozygous for the *A* allele"). An individual whose genotype is *BO* at that locus would be called a heterozygote.

In summary, the genotype of an individual at a locus describes the allelic forms of DNA that an individual possesses. One major difficulty with determining someone's genotype is that it is in many cases extremely difficult or impossible to perform the molecular analysis necessary to determine the DNA sequence that exists on that individual's chromosomes at that locus. Thus, directly observing genotypes is often not possible or at the least, not practical.

But what is often much simpler to observe is what is called a *phenotype*. An individual's phenotype is an outwardly visible or otherwise apparent characteristic of the individual. An observed phenotype may be the result of one or possibly many underlying genes, as well as be affected by the environment.

For example, a human's eye color can be observed quite easily. So we would describe someone's eye color phenotype as either blue, green, brown, hazel, etc. What is more difficult is to describe that individual's eye color *genotype*. For one, the exact biological and genetic (and environmental) basis for eye color is not completely understood. Second, the genetic understanding we do have for eye color determination is a complex interaction among at least three genes. So it would take a potentially difficult molecular analysis to describe the genotype.

Example 1.3: **ABO Blood group**: Although there are six genotypes possible at the ABO locus, there are only four observed phenotypes in individuals. These four are the typical ways that a person's blood type is denoted:

Phenotype	Corresponding Genotype(s)
A	AA, AO
B	BB, BO
AB	AB
O	OO

Furthermore, it is a simple matter of analyzing a person's blood-workup to determine their blood type (the phenotype). We can see from the table that knowing your blood type may also let you directly infer what alleles you must carry in your DNA at the ABO locus. For example, if you are blood type AB, it must be that your genotype is AB, meaning you have one A allele and one B allele. But if your blood type is A, you could either be genotype AA or AO. The reason for this has to do with the relationship between these alleles, which will discuss further below.

■

The different possible allelic forms of a locus can interact in various ways in an individual. An allele is called a *recessive* allele with respect to another if its phenotypic effect is suppressed in individuals heterozygous for those two alleles. Therefore, a recessive allele can only be expressed phenotypically in individuals who are homozygous for it.

In the above paragraph, the allele that "suppressed" the recessive allele is called a *dominant* allele with respect to the other. Two alleles that interact in this way are said to have a *dominant-recessive relationship*.

The other relationship two alleles can have is a *codominant* relationship. In this case, both are expressed phenotypically in an individual who is heterozygous for those alleles.

Example 1.4: **ABO Blood group**: This locus continues to present a nice example for us because among its three alleles are all the various relationships mentioned above:

- The O allele is recessive to A and B. An individual has to have two copies of O (genotype: OO) for its effect to be evident (and therefore have the O blood type phenotype).
- The A allele dominant to O. An individual with an AO genotype shows no phenotypic effect from having the O allele.
- Same as above for the B allele with respect to the O allele.
- The A and B alleles are codominant. An individual with the AB genotype has the effect of both alleles apparent in their phenotype (which is called the AB blood type).

■

Example 1.5: **Albinism**: Lack of pigment in humans results in an individual having albinism (i.e., an albino). A particular gene is associated with having albinism. There are two possible alleles, which we'll refer to as A and a. Allele A is dominant to allele a (this, by the way, is a very standard notation for alleles with a dominant-recessive relationship). A person only has albinism if they have genotype aa. AA and Aa genotypes are normal with respect to pigment. Traits such as albinism are called *recessive traits*.

■

Example 1.6: **Other Traits**: Most traits in individuals are not simply determined by the alleles at one gene. Many traits are expressed as a result of an individual's genotype for two, three, or possibly many genes. As we briefly mentioned earlier, eye color is known to be (mostly) a result of the interaction of three genes, but possibly more. In this more complex system, brown is dominant to green, which is dominant to blue, which is recessive to both.

■

Let's finish this section by reviewing and presenting some further notation. In many analyses, we are interested in two or more loci at the same time. A general notation for the loci in an analysis will be to use consecutive capital letters such as **A**, **B**, **C**, etc. The allelic forms of a locus will generally be notated using the letter representing the locus subscripted by appropriate integers (such as A_1, A_2, A_3, or B_1, B_2). This type of subscripted notation will particularly be used for alleles which have a codominant relationship.

For genes which have two alleles with a dominant-recessive relationship, our standard notation will be to name the dominant allele using the same capital letter as used to notate the locus. The recessive allele will be named using the lower-case version of that letter. Examples are alleles A and a for locus **A**, or alleles B and b for a locus **B**.

1.4 Genetic Inheritance

We will now review the basics of how DNA is inherited from one generation to the next. First recall that each individual has two copies of each chromosome (the two homologous chromosomes) in each diploid cell. We'll ignore for now situations such as the sex chromosomes in males. One copy of each chromosome came from the individual's mother and one from the individual's father. So, at any particular locus, one allele possessed by an individual is one of the two possessed by the mother, and the second allele for the individual is one of the two possessed by the father.

Example 1.7: **Various examples:**

ABO Blood group mating example:

Mother's Genotype:	AO
Father's Genotype:	BO
Offspring Potential Genotypes:	AB,AO,BO,OO
Offspring Cannot Be:	AA, BB

Albinism: For this locus, two "normal" individuals who mate can only have an albino offspring if both were heterozygote Aa. They could then both pass on the a allele to their offspring, making the offspring an aa genotype and therefore albino.

An AA mating with an Aa cannot produce albino offspring because offspring will always get at least one A allele from the first parent. Aa individuals are called *carriers* for recessive traits because although they don't have the trait, they carry the gene for it and could potentially pass it on to their offspring.

■

The process of inheritance begins with a cellular process called *meiosis*. Meiosis creates four haploid cells (called gametes) from one diploid cell. Understanding meiosis will be key to many genetic analyses we'll discuss. A simplified version of what happens in meiosis is:

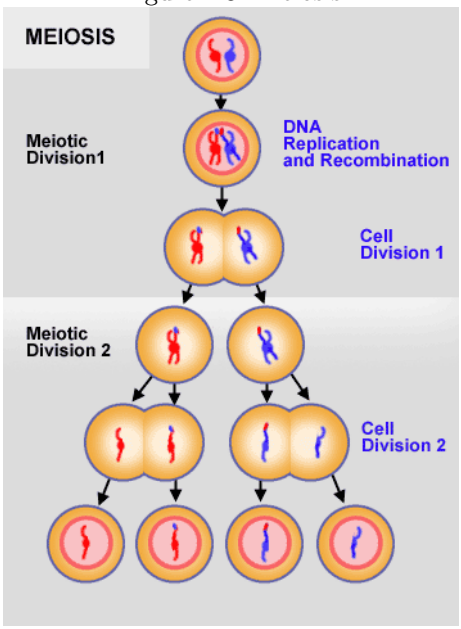
1. Start with a diploid cell containing $2n$ chromosomes, where n is the species' chromosome number.
2. Each chromosome replicates itself to form a *sister chromatid* which is an exact copy of the original. At this point, we can think of the cell as having $4n$ chromosomes present.
3. The original homologous chromosomes attach to each other, which forms *chiasma* (pl. chiasmata) between them and produces *crossover* events, thereby exchanging genetic material (more on this later).
4. A much simplified version of what happens next is that four new haploid cells are formed, with one of the four copies of each chromosome going to each new cell in a random fashion. The fact that this segregation of chromosomes happens in a random fashion is very important for our forthcoming analyses.

Below is a link to an animated movie that shows meiosis in detail.

http://fig.cox.miami.edu/Faculty/Dana/meiosis.mov

The products of meiosis are called *gametes*. They are haploid cells that have half the genetic information that their parent cell possessed and are involved in sexual reproduction. When two gametes join up in a mating process, for instance human male sperm and female and egg gametes, the genetic information is linked together to make a diploid zygote and eventually a new individual with the combined genetic information of its maternal and paternal parents. Figure 1.8 shows an overview of the meiosis process.

Figure 1.8: Meiosis



Let's walk through a few examples of matings and offspring to better understand genetic inheritance and meiosis. We'll also discuss some simple probability concepts by introducing the idea of expected genotypic ratios.

Let's first look at a simple example where we are studying inheritance at one locus. We can use the albinism gene as a basis for our discussion. Let's refer to the locus as **A**, and refer to the two possible alleles as A and a just as before (a is the recessive allele causing albinism).

Now, let's say that an individual with genotype Aa (female) mates with an individual with genotype AA (male). In notation, this mating is often written as $Aa \times AA$, where the $\times$ is read as "cross". What are the possible offspring genotypes, and in what proportions would we expect them to occur?

Let's first walk through this problem by thinking about how meiosis works. Meiosis in the Aa individual will produce gametes of which $1/2$ have the A allele, and $1/2$ have the a allele. In other words, $1/2$ of this woman's eggs carry A and $1/2$ carry a.

Meiosis in the AA individual will produce gametes all of which contain the A allele. All sperm of this man carry the A allele.

When mating occurs between the two, a union is formed between a random egg from the first individual and a random sperm from the second individual. This union forms a cell (a zygote) which is diploid and carries 46 total chromosomes. With regard to locus **A**, it should be easy to see that there is a $1/2$ chance that this zygote will have genotype AA, since the male will always contribute an A allele and the female has a $1/2$

chance of contributing an A allele. Similarly, there is also a 1/2 chance that the offspring will have genotype Aa and be a carrier just like his/her mother.

Such an analysis can be done much more quickly using a table like that in Table 1.5. The table is constructed by listing the gametes carried by one parent as the columns and listing the gametes carried by the other parent as the rows. The percent of each type of gamete for each parent is also listed. Then each cell in the table is filled in by writing down the resulting offspring genotype for that particular combination of gametes, and the product of the two associated percentages.

The cells of the table then give all possible offspring genotypes along with the percentage of offspring we expect to have that genotype.

Let's look at a second example of a mating using the albinism gene. This time, two carriers Aa mate. Table 1.6 shows the results. First notice that since this was slightly more complicated than the last example, the results in the table need to be summarized. Specifically, the offspring genotype Aa occurs twice in the table, so we need to combine those cells and add up the percentages to get the total percentage for Aa. Notice that there is a 1/4 chance that an offspring of this mating will have albinism (i.e., have genotype aa), and a 1/2 chance that an offspring will be a carrier.

For one last one locus example, let's return to the ABO blood group locus and consider the mating of an AO genotype with a BO genotype. Table 1.7 shows the results. An offspring of this mating has an equal chance (1/4) to be any of the four blood type phenotypes.
We will also look at an example where we are interested in two loci that are on different chromosomes. This will be covered in class.

There is another very important event that occurs during meiosis that we have only glossed over so far. It is called *crossing over* and occurs between pairs of homologous chromosomes. The overall result is that the two homologous chromosomes will exchange genetic material, causing alleles from one chromosome to be integrated into the other, and vice versa.

Further, it is possible that two crossing-overs can happen during meiosis on the same chromosome. This can complicate matters because the effect of one crossing over (in terms of the recombinant gametes produced) can then be negated by the second crossing over. This is the reason that most linkage analyses are conducted in very small regions of a chromosome, so that the probability of a second crossing over is very unlikely. Three or more can occur as well, with smaller likelihoods.

More details and examples on crossing over and recombination will be given during lecture. Also refer to Figures 1.9 and 1.10.

Loci that are on the same chromosome are often called *linked loci*. If the loci represent genes, they are, of course, more specifically called *linked genes*. You can think of this phrase as referring to the fact that the

Figure 1.9: Crossing Over

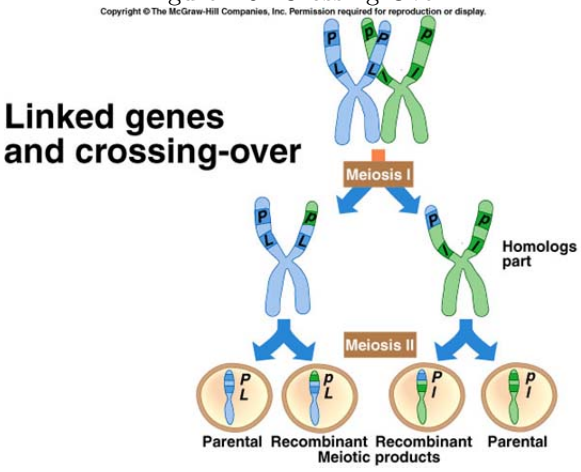
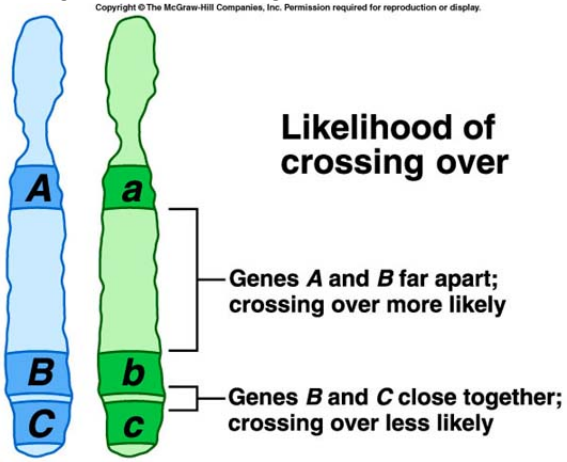


Figure 1.10: Crossing Over and Linked Loci



loci are physically "linked" to each other on the same strand of DNA.

An important quantity that we will introduce briefly now, but spend much more time discussing later in the course is called the *recombination fraction*. A recombination fraction is always measured with regard to two loci. In a sense, it can be thought of as a measure of how close the two loci physically are to each other on a chromosome.

More specifically, the notation we will use is $r_{\mathbf{AB}}$ where **A** and **B** are the two loci in question. The formal definition for this quantity is that $r_{\mathbf{AB}}$ is the proportion of gametes that are recombinant with regard to the two loci in some number of observed meioses events. Theoretically, it is the case that $r_{\mathbf{AB}} \in [0, 0.5]$ since (as we saw in class) one-half of the gametes resulting from a meiosis are guaranteed to be non-recombinant. In practice, however, it is possible to measure $r_{\mathbf{AB}}$ to be larger than 0.5, but we typically rewrite it as 0.5 if this happens.

A value at or near 0 suggests that the two loci are very near each other physically on the loci, since there are very few crossing overs occurring between them. A value near 0.5 can be interpreted as the loci are either far apart on the same chromosome, or on totally different chromosomes.

If you followed our in-class demonstration of crossing-over closely, you should also notice that $r_{\mathbf{AB}}$ will equal one-half of the crossover rate $c_{\mathbf{AB}}$ between the loci, where $c_{\mathbf{AB}}$ is defined as the percentage of meioses where an odd number of crossovers occur between **A** and **B**.

More mating examples where we consider two loci linked on the same chromosome will be covered in class.

One use of estimating values of $r_{\mathbf{AB}}$ for many pairs of loci is to develop a *linkage map* of a chromosome. This is a map of the relative positions of genetic loci on a chromosome, determined by estimating recombination percentages. Distance between loci is measured in units called *centimorgans* (cM). One centimorgan defined to be equal to a 1% chance that a marker at one genetic locus will be separated from a marker at a second locus due to crossing over in a single generation. In human beings, 1 centimorgan is equivalent, on average, to 1 million base pairs (although it differs in different regions of the genome).

Linkage maps will be discussed in more detail later in the course. For now, you should know that they have been used extensively to help locate (i.e., pinpoint the physical location) of genes that cause diseases in humans, as well as genes that play other important roles.

Linkage Exercise: Corn

(courtesy of http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/L/Linkage.html)

Start with two different strains of corn (maize).

- one that is homozygous for two traits yellow kernels (C,C) which are filled with endosperm causing the kernels to be smooth (S,S).
- a second that is homozygous for colorless kernels (c,c) that are wrinkled because their endosperm is shrunken (s,s).
- C is dominant to c, and S is dominant to s.

When these two strains are crossed, the kernels produced (the F1 generation) are all yellow and smooth (all will be CcSs) because of the dominance relationship.

Now, mate these F1 corn with the second strain above (the homozygous recessive strain ccss). Such a mating is called a test cross because it exposes the genotype of all the gametes of the strain being evaluated.

Analyze what would happen in this cross if

(1) the two genes were on different chromosomes, and

(2) the two genes were so closely linked that there were no chance of crossing over between them.

Then compare to the actual results (to be discussed in class).

Table 1.3: Common human syndromes caused by structural abnormalities in chromosomes

Monosomy: individual has only one of a chromosome pair

- **Turner's syndrome** - XO - Commonest of all abnormalities of chromosome number - only 2% of zygotes with XO survive to term. This individual is female with underdeveloped sex organs, webbed neck, shorter than average height, with heart and circulatory defects common.

Trisomy - individual has three (or more) of a chromosome pair

- **Superfemale** - XXX, XXXX, XXXXX - Mental retardation is common, as are body structure derangements, as well as reduced fertility, especially in XXXX and XXXXX.
- **Klinefelter's syndrome** - XXY, XXXY, XXXXY - Body structure derangements include feminization and elongated extremities. These males are typically sterile.
- **Supermale** - XYY - Little change in outward appearance is evident.
- **Down's syndrome** - Trisomy 21 (3 copies of Chromosome #21)- Mental retardation, shorter than normal stature and increased risk for leukemia are some of the manifestations of this condition. Down's affects 1 in 500 births, and is detected in 20% of the spontaneously aborted fetuses.
- **Patau's syndrome** - Trisomy 13 - Cleft lip and palate, cerebral, ocular and cardiovascular defects are evident. Life expectancy is less than 3 months.
- **Edward's syndrome** - Trisomy 18 - Life expectancy is less than 6 months.

Deletions - large areas of a chromosome missing

- **cri-du-chat** - deletion on long arm of chromosome 5 - Phenotype evidenced by small head, mental retardation and catlike cry.
- **Duchenne muscular dystrophy** - deletion on short arm of X - Manifested as a muscular degenerative disease causing death generally by age 18.
- **Lesch-Nyhan syndrome** - deletion of HGPRT locus on long arm of X - Self-mutilation is common in Lesch-Nyhan sufferers.

Deletion and translocation - missing and moved to another chromosome

- **Philadelphia chromosome** - deletion of half of the long arm of chromosome 22 - deleted material may be translocated to long arm of chromosome 9 - found only in leukocytes of patients with myeloid leukemia.
- **Retinoblastoma** - deletion of material of long arm of chromosome 13 - Manifested as childhood cancer of the retina.

Table 1.4: Estimated number of genes for various species

Organism	**Estimated No. Genes**
Bacterium	3,025
Yeast	6,225
Fruit fly	12,000
Plant	20,000
Human	25,000

Table 1.5: Computing genotype ratios for for $Aa \times AA$

	Parent 1 Gametes	
	A	a
Parent 2 Gametes	$\frac{1}{2}$	$\frac{1}{2}$
A	AA	Aa
1	$\frac{1}{2}$	$\frac{1}{2}$

Table 1.6: Computing genotype ratios for $Aa \times Aa$

	Parent 1 Gametes	
	A	a
Parent 2 Gametes	$\frac{1}{2}$	$\frac{1}{2}$
A	AA	Aa
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
a	Aa	aa
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

Table 1.7: Computing genotype ratios for $AO \times BO$

	Parent 1 Gametes	
	A	O
Parent 2 Gametes	$\frac{1}{2}$	$\frac{1}{2}$
B	AB	BO
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
O	AO	OO
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

Chapter 2

Probability and Statistics Review

This chapter should serve as a good review of basic concepts in probability and statistics. The material in most sections should not be new to you, but there are a few topics which you mostly likely have not seen in a course before. We will cover these in more depth in the course.

2.1 Introduction to Random Variables

When an experiment is conducted, many characteristics are often measured during and following the experiment. The exact numerical value obtained for any particular measurement will depend on many factors such as environmental factors, experimental conditions, or just unexplained factors that always result from taking measurements in real situations. We will call such a measurement a random variable.

A *random variable* is a variable whose value, when measured during or following an experiment, may be one of two or more possible numeric values. The particular value measured depends on various random factors. Capital letters like X and Y will be used to represent a random variable.

There are two key parts to this definition. First, the value measured for a random variable must be numeric. Often, the measured value is restricted to some domain, such as the positive integers or the real numbers between zero and one, as we'll see in examples below.

Second is the concept of *random*. The value of the random variable that results from an experiment is not known in advance. All we know is that it will be some value from a domain of possible values. Some values are more likely to result than others. The relative likelihoods of the various possible values in the domain are what we define when we set up our probability model for the experiment, as we will discuss in this chapter. The domain of possible values of a random variable (also to be called the *set of possible values* or just *the possible values*) will be denoted with the script version of the letter that symbolizes the random variable (such as $\mathcal{X}$ for X or $\mathcal{Y}$ for Y).

There are two basic types of random variables: *discrete* and *continuous*. A discrete random variable is a random variable whose domain of possible values is finite or countably infinite. Typically, the possible values for a discrete random variable will be all non-negative integers, all positive integers, or some finite set of integers. Below are some simple examples of discrete random variables along with their possible values:

- X = Result from one roll of a die. $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$.
- X = Number of coin tosses it takes to get the first heads. $\mathfrak{X} = \{1, 2, 3, \ldots, \infty\}$.
- X = Number of crossovers that occur between two loci during a meiosis. $\mathcal{X} = \{0, 1, \ldots, N\}$ where N might be some known maximum number of possible crossovers that could occur depending on the two loci and other biological factors.
- X = Number of matching bases when comparing two strands of DNA each of length N. $\mathcal{X} = \{0, 1, \ldots, N\}$.
- X = Number of LEU amino acids in a randomly selected human protein product. $\mathcal{X} = \{0, 1, \ldots, N\}$. Again, the maximum possible value N might be defined as the size of the longest known protein.

From these examples, we can see that discrete random variables usually result from counting the number of times something occurs during an experimental period.

A continuous random variable is a random variable whose set of possible values is uncountably infinite, such as an interval of the real line. Below are some simple examples of continuous random variables along with their possible values:

- X =Recombination percentage between two loci measured from an animal mating experiment. $\mathcal{X} = [0, .5]$.
- X =Time until a certain cellular molecule degrades. $\mathcal{X} = (0, \infty)$.
- X =Molecular weight of a randomly selected RNA molecule. $\mathcal{X} = (0, M)$ where M is some constrained upper bound for such a quantity.

As we can see, continuous random variables are often variables that measure quantities such as times, lengths, weights, and percentages.

Let us relate this discussion of random variables back to our discussion of probability back in Chapter 3. We still are conducting experiments, and we are still interested in the probability of various outcomes. However, now all interesting events for which we want to calculate probabilities will be referred to using the notation of random variables.

For example, if we let X = Number of matching DNA bases across two aligned DNA strands of length N, we might be interested in events such as finding that more than half of the bases match, or the event that all of the bases match. Written in terms of the random variable X, we would write these probability statements as $\mathrm{P}(X > N/2)$ and $\mathrm{P}(X = N)$, respectively. Notice that the events inside of the probability statements are each denoted by mathematical equalities or inequalities involving X, as opposed to a list of outcomes.

To provide an example for a continuous random variable, let X = recombination percentage between two loci. We may want to calculate probabilities such as $\mathrm{P}(X < 0.2)$ or $\mathrm{P}(X > .45)$. Again, the events in question are written in terms of the random variable. In this case, since X is continuous, the range of values of X covered by these events is a continuous range of real numbers.

To conclude this section, we note that the two types of random variables are very different in nature when it comes to defining and working with probability models, so they must be dealt with separately. In the sections that follow, we will introduce discrete random variables first, then continuous random variables. We will cover a number of special cases of each of these types that are important in probability modeling.

2.2 Discrete Random Variables

In this section, we will begin our discussion of random variables by focusing on the discrete type.

2.2.1 Probability Distribution

Any discrete random variable has an associated *probability distribution*. This is also called its *probability mass function* or pmf. The probability distribution is a list of the possible values of the random variable along with each of their probabilities. In more complex cases, this information is more easily displayed by use of a formula instead of a "list".

For a probability distribution to be valid, it must satisfy two properties. These properties are a direct result of the axioms of probability we saw in Chapter 3. They are

$$\sum_{x \in \mathcal{X}} \mathrm{P}(X = x) = 1$$

and

$$\mathrm{P}(X = x) \geq 0 \quad \forall\, x \in \mathcal{X}.$$

The first states that the probabilities assigned to each value in the domain of X must sum to one. This is equivalent to our probability axiom which states that the probability of the sample space must be exactly one.

The second property above simply restates that probabilities must be non-negative numbers. As we saw in Chapter 3, this combined with the first property above leads to the further restriction that all probabilities are in fact between zero and one inclusive.

Example 2.1: Let X=number of crossover events during a particular meiosis along a certain length of DNA. We will specify a probability distribution for this random variable X. Say $\mathcal{X} = \{0, 1, 2, 3, 4, 5\}$. Then we might write the probability distribution as:

$$\mathrm{P}(X = x) = \begin{cases} .5 & x = 0 \\ .3 & x = 1 \\ .15 & x = 2 \\ .03 & x = 3 \\ .01 & x = 4 \\ .01 & x = 5 \end{cases}$$

Note the probabilities add to one, so this is a valid probability distribution.

■

It is important to keep in mind that a probability distribution, as in Example 2.2.1, is also, in fact, a function. In the example, the left hand side $\mathrm{P}(X = x)$ is often written $p(x)$ or $p_X(x)$ to make it clear that we are defining a function of x (recall that this is also called a probability mass function). The right hand side defines the function; for various real numbers that we might plug into the function, it tells us what the function returns. In other words, $p(0) = 0.5$ and $p(3) = .03$, for example.

Also, the probability distribution as given in the example is not complete as written. For example, it doesn't specify what the function returns if we plug, say, $x = 2.6$ or $x = -32$ into the function p. In fact, to be complete, the probability distribution should also have one more case that states $\mathrm{P}(X = x) = 0$ for all other real number x. This information is commonly left out of the definition of a pmf, and it is assumed that any values for which the function has not been defined have probability zero.

Let us continue with one further note about this example. Why did we use those particular probabilities in our definition of the probability distribution? At this early stage of our discussion of random variables, it will suffice to consider that they were determined based on our experience with collecting this type of data and with this type of experiment. The given probabilities may or may not be validated by data we ultimately collect.

One word on notation is necessary before we move on. Notice that we have used two versions of the letter "X" in the example - a lowercase and uppercase version of the letter. There is a fine but important distinction between the two. The uppercase version of the letter (i.e., X) denotes the random variable itself. In other words, X is the random variable that represents the number of crossovers. The lowercase version (i.e., x) represents a particular possible value of the random variable. So the statement $\mathrm{P}(X = x)$ can be read as "the probability that the random variable X will equal the particular value x." Different values plugged in for x lead to different probability statements, such as $\mathrm{P}(X = 1)$ or $\mathrm{P}(X = 3)$.

We can see that the pmf of a discrete random variable provides us with the most basic information about the random variable. Namely, it gives us the likelihood of any particular value being observed when the

experiment is conducted. When we write down the probability distribution, we are making a statement about what values we believe are more or less likely to result. In the crossovers example, we believe the value 0 is the most likely to be observed (i.e., no crossovers), while 1, 2, or more crossovers are less and less likely to occur. In fact, we the model states that in 50% of such experiments, there will be no crossovers.

The pmf allows us to calculate any probability question we may ask. For example, the probability that there will be 2 or more crossovers during the meiosis is

$$\mathrm{P}(X \geq 2) = \mathrm{P}(X = 2) + \mathrm{P}(X = 3) + \mathrm{P}(X = 4) + \mathrm{P}(X = 5) = 0.2$$

and the probability that there will be an even number of crossovers is

$$\mathrm{P}(X \text{ is even}) = \mathrm{P}(X = 0) + \mathrm{P}(X = 2) + \mathrm{P}(X = 4) = 0.66.$$

An interpretation we can attach to this last statement is "if we were to observe such a meiotic event over and over again many times, we would observe an even number of crossovers between the loci 66% of the time."

2.2.2 Cumulative Distribution Function

Another special function associated with any discrete random variable is its *cumulative distribution function* or cdf. This function is defined for any real number x as the probability that the random variable will be less than or equal to x. The function notation F is often used to represent the cdf of a random variable. By definition, we can write

$$\mathrm{F}_X(x) = \mathrm{P}(X \leq x).$$

Note that we have subscripted the F function notation with the letter of the random variable. This makes it clear which random variable's cdf we are talking about.

Based on the crossover model in Example 2.2.1, we would have

$$\mathrm{F}(x) = \begin{cases} 0 & x < 0 \\ .5 & 0 \leq x < 1 \\ .8 & 1 \leq x < 2 \\ .95 & 2 \leq x < 3 \\ .98 & 3 \leq x < 4 \\ .99 & 4 \leq x < 5 \\ 1 & x \geq 5 \end{cases}$$

Be sure to notice that this function satisfies our definition for a cdf. Plug in any real number x, and it returns $\mathrm{P}(X \leq x)$. This includes values not necessarily in random variable's domain, such as -1, 3.4, or 7.2

in this example.

Notice that the pmf and cdf both provide us with the same information. One can always be derived from the other. Which we use is often a matter of convenience.

2.3 Continuous Random Variables

We have first discussed discrete random variables and some introductory applications of such random variables. In order to progress into more topics, we need to understand continuous random variables as well.

As we saw earlier, the key difference for a continuous random variable is that its set of possible values is uncountably infinite, or in other words, is a range of real numbers. That range could begin at $-\infty$, or end at $+\infty$, or both. Clearly, random variables that measure distance or time will tend to fall into this category, as well as other quantities that can be measured with fine precision. Actually, sometimes even a truly discrete random variable may be modeled as a continuous random variable if it has a very large number of possible values. For example, if $X \sim \text{Bin}(200, p)$, then the range of values for X is $[0, \ldots, 200]$ which is often modeled continuously (as we'll see later).

Some examples of continuous random variables we might come across are (some were first mentioned earlier):

X =Time until a protein degrades in the cell: $\mathcal{X} = (0, \infty)$

X =Distance between two loci on a chromosome: $\mathcal{X} = (0, \text{Chromosome Length})$

X =Recombination fraction between two loci: $\mathcal{X} = [0, .5]$

X =Log of Relative gene expression level for a certain gene: $\mathcal{X} = (-\infty, \infty)$

X =Amount of a certain antibody produced by a random individual: $\mathcal{X} = (0, \infty)$

2.3.1 General Details About a Continuous Random Variable

For a discrete random variable, its pmf or probability mass function defines all that we need to know about the random variable. We use it to understand the distribution of possible values, directly calculate probabilities, and compute the mean and variance of the random variable.

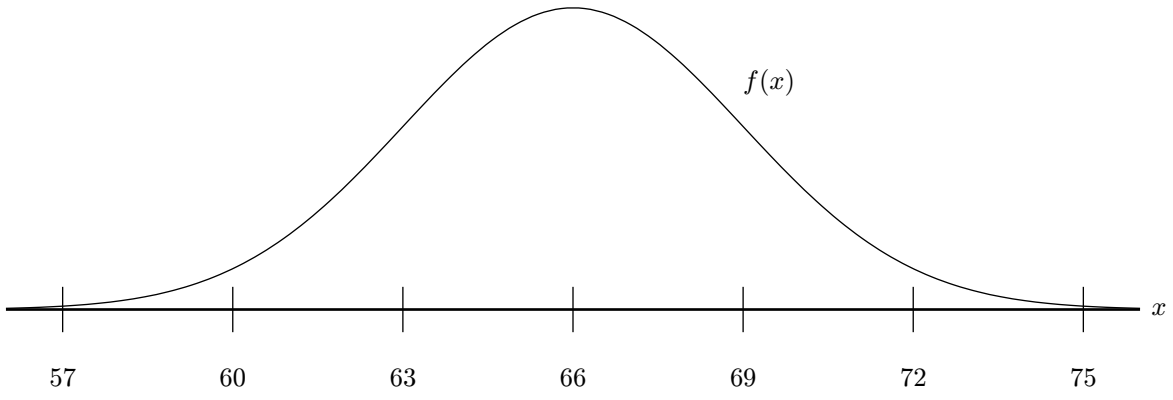
A continuous random variable has an analogous function associated with it called its *probability density function* or *pdf*. There is a very important difference between the two, which the difference in terminology tries to get across. A pmf for a discrete random variable is defined (with positive probabilities) only for a finite or countably infinite set of possible values - typically integers. A pdf for a continuous random variable is defined for all real numbers in the range of the random variable. It is a continuous function, more typical of functions encountered in a calculus course. For this reason, we will use the notation $f(x)$ to denote a pdf.

To be clear that it is the pdf of the random variable X, we will often subscript f with X, as in $f_X(x)$.

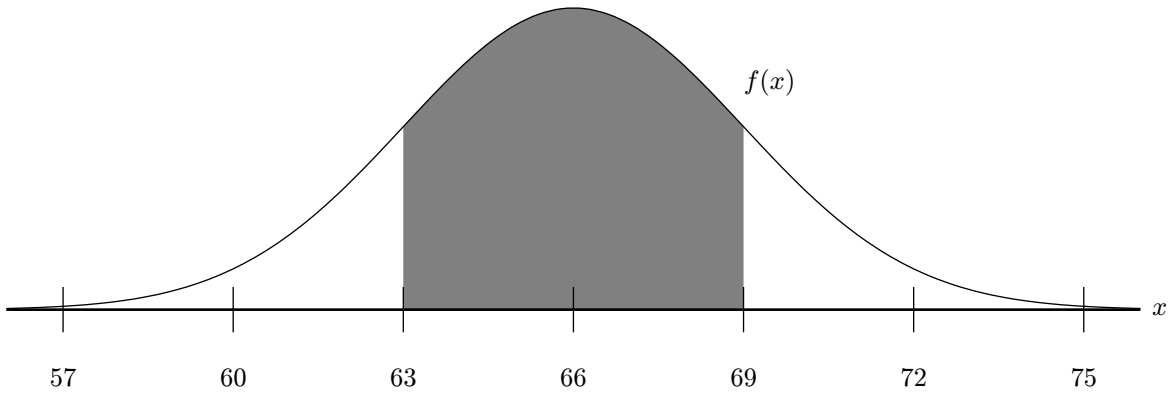
Now, the choice of such function to use for a certain random variable has everything to do with how we want to model probabilities of various sets of values occurring. The overriding concept here for a continuous random variable is that **AREA=PROBABILITY**. More specifically, the area under the pdf curve between points a and b is the same as the probability that the random variable will have a value between a and b. Since the computation of an area is an integration problem in calculus, we can write this mathematically as

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx.$$

To view this from a graphical standpoint, let's consider the random variable X=Height of a randomly selected adult (in inches). Clearly, we can consider X as a continuous random variable, and we need to decide how to model such probabilities. In other words, we need to select a function $f(x)$ that does a reasonably good job of representing the true distribution of heights in adults. Suppose we choose a certain function $f(x)$. Without worrying for now about the actual functional form of this f we selected, let's plot it against x and take a look at the resulting picture (by the way, let's also not worry for now about how we go about choosing the f we choose; we'll discuss that later):



Remembering that area and probability are the same, this plot suggests that a high percentage of adults are between, for example, 63 and 69 inches tall (since the curve has a lot of area under it between those points). Precisely, that percentage is $\mathrm{P}(63 \leq X \leq 69) = \int_{63}^{69} f(x)dx$, or viewed graphically as the area shaded in below:



It also suggests that a very small percentage of adults are taller than 75 inches, or shorter than 57 inches (since the curve has very little area under it beyond to the right and left of those points respectively). We could get more detailed and describe the probability, or percentage of adults, that have a height between any two values by observing the area under the curve between those values. All-in-all, we can probably agree that this particular $f(x)$ does a reasonable job of describing the heights of adults. Its *shape* seems like an accurate description of real-life.

2.3.2 Properties of the density function $f(x)$

It is probably clear that the $f(x)$ we decide to use as the pdf for X cannot be just any function at all. There are a couple of key restrictions that it must adhere to if our concept of Area=Probability is going to make sense. These restrictions are:

1. $f(x) \geq 0$ for every x
2. $\int_{-\infty}^{\infty} f(x)dx = 1$

Both are necessary if we are to be sure that areas we calculate will adhere to rules of probability. The first is needed because if we allow $f(x)$ to drop below the x-axis, then certain area calculations, and hence probability calculations, would come out negative. So, the curve $f(x)$ must always be at or above the x-axis.

The second is equivalent to the fact that the probability of *something* occurring is 1, or 100%. If we allow a pdf to be defined such that the entire area under the curve is more than 1, for example, that would lead to certain probability calculations being more than 1, which is invalid. If we allow it to be such that the entire area under the curve is less than 1, then an outcome may occur that we haven't accounted for in our model.

These restrictions are important, but don't limit our choices for $f(x)$ very much. There are still an infinite number of possible functions that meet these requirements. In particular, consider a function $g(x)$ whose

area under its curve is some number $A \neq 1$ and which is always non-negative. Then the function $f(x) = \frac{g(x)}{A}$ will be a valid pdf, and will have the same basic shape as $g(x)$, just scaled by A. So for any function $g(x)$ that doesn't have an area of 1, we can turn it into a valid pdf just by dividing by that area.

By the way, it is the second fact that makes it often unnecessary for us to label the units on the vertical axis when we plot $f(x)$. We know that the units must be such that the area under the plotted curve must equal one. And, usually the exact height of the curve at any particular point is not helpful to know. We just need to know area under the curve between two points.

One other thing to note is that, since the area under a single point is precisely zero, then $\mathrm{P}(X = a) = 0$ for any real number a. That is, the calculated probability that a continuous random variable will equal a particular real number is zero. We can only get positive probabilities if we make the calculation over a contiguous range of real numbers such as 63 to 69 or 72 to ∞ in the above examples.

This fact also makes a "less than sign" the same as a "less than or equal to sign" (and same for "greater than") in probability calculations. In other words $\mathrm{P}(X < 64) = \mathrm{P}(X \leq 64)$. We will still strive to use the inequality sign appropriate to the situation.

2.3.3 Cumulative Distribution Function

The cumulative distribution function (cdf) for a continuous random variable is the continuous function defined by

$$F_X(x) = \mathrm{P}(X \leq x)$$

for any real number x. Since the function F is defined as a probability, it only returns values in the range [0,1]. Plugging any real number x into $F_X(x)$ returns the probability that the random variable X will have a value less than or equal to the number x.

Note that $F_X(x)$ and $f_X(x)$ are closely related by this definition:

$$F_X(x) = \int_{t=-\infty}^{t=x} f_X(t)dt$$

where t has been used as the dummy variable of integration to be perfectly correct from a calculus standpoint. From the other direction, we have

$$f_X(x) = \frac{dF_X(x)}{dx}$$

and so it is clear that given one of these functions, we can determine the other (assuming the integral or derivative is not too difficult!).

2.4 Expected Value and Variance

A pmf gives all the information we need about a discrete random variable, while a pdf does the same for a continuous random variable. We can calculate the probability of any event we want from them. However, we often like to summarize the distribution more concisely, such as for the purpose of comparing one random variable to another. Two common calculations that are made for a random variable are called the *mean* or *expected value* of the random variable, and the *variance* of the random variable.

The mean of a random variable is defined as the average value that would be observed for the random variable if it could be observed over and over again an infinite number of times (or more practically you might think of this as just many, many times). It is defined as

$$\mu_X = \sum x\mathrm{P}(X = x) \tag{2.1}$$

for a discrete random variable, and

$$\mu_X = \int xf(x)dx \tag{2.2}$$

for a continuous random variable. The sum or integral is taken over all possible values of X, that is, over $\mathcal{X}$. In either case, it is reasonable to interpret this calculation as taking the weighted average of the possible values of X, where the weights are the probabilities.

The variance of a random variable is a measure of how spread out its possible values are. A random variable will have a small variance if its possible values all fall in a small, tight range with high probability. It will have a large variance if its possible values are very spread out over a wide range and there is a reasonable chance than any of those values could be observed. The variance calculation is

$$\sigma_X^2 = \sum x^2\mathrm{P}(X = x) - \mu_X^2. \tag{2.3}$$

For a continuous random variable, replace summation with integration and "P" with "f". Another equivalent formula for the variance which is sometimes easier to deal with is

$$\sigma_X^2 = \mu_{X^2} - (\mu_X)^2.$$

A calculation related to the variance is called the *standard deviation* of a random variable. It is simply defined as the positive square root of the variance. This is actually used more often in practice than the variance because the units of the standard deviation calculation are the same as the units of the original variable, while the units of the variance are the square of the units of the original variable.

$$\sigma_X = \sqrt{\sigma_X^2}$$

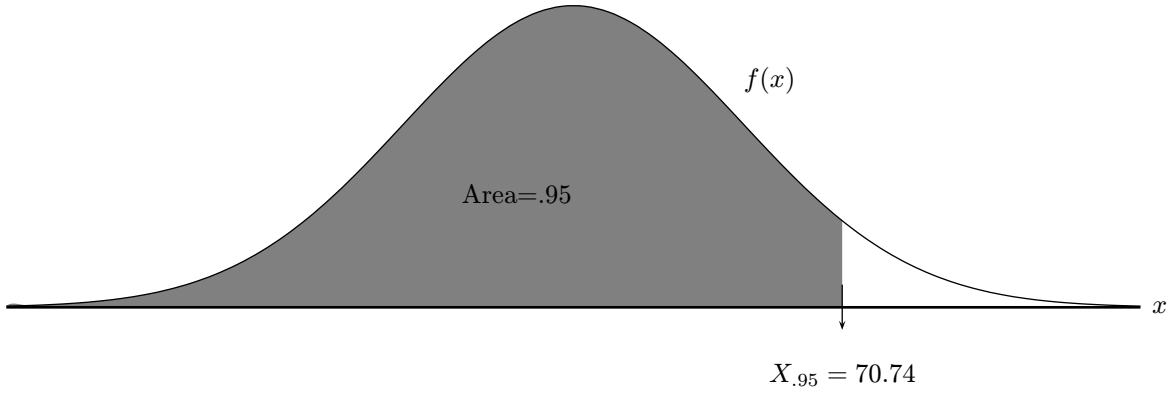
2.5 Percentiles of a Random Variable

In many situations, we want an answer to questions like, "what value of X is such that 90% of all values are below it?" The answer to this is called the *90th percentile of* X. Of course, we can choose to ask for any percentile, not just the 90th.

Answering such a question is accomplished by looking the inverse of the cdf. The cdf is used to answer questions like: "given a possible value of X, return the probability of being less than or equal to that value." This is the inverse of the percentile question: "given a probability (i.e., percentile), return the possible value that falls such that the given percentage of all possible values are less than or equal to it."

We will notate this inverse function as $F_X^{-1}(\alpha)$ where alpha is the percentile we want ($0 < \alpha < 1$). So, for example, $F_X^{-1}(.63)$ is the 63rd percentile of the distribution of X, or the value that is such that 63% of the area under the curve is to its left. Or, since this notation can be cumbersome, instead of writing $F_X^{-1}(\alpha)$, we might just write X_α in general, or $X_{.63}$ as a specific example.

As an example, it turns out in our height example that 70.74 inches is the 95th percentile of height according to that probability model. 95% of all adults have a height less than or equal to 70.74 inches, and so 5% are taller than 70.74 inches. In notation, this says that $X_{.95} = F_X^{-1}(.95) = 70.74$, or $\mathrm{P}(X \leq 70.74) = .95$, or in a picture



It turns out that it is fairly easy to find percentiles if we have access to software packages, even simple spreadsheets like Excel. So, this is how we will rely on making such calculations.

2.6 Common Discrete Distributions

The previous discussion was a very general discussion of discrete random variables, and the definitions and ideas apply for any such random variable. However, in real applications, the probability model (i.e. distribution) we use for a random variable often falls into one of certain special cases depending on how that random variable is measured and the way the experiment is conducted. There are very many such special cases, but a few tend to occur often in real genetic applications.

2.6.1 Bernoulli Random Variable

A Bernoulli random variable is perhaps the simplest of all possible discrete random variables. A random variable is called a Bernoulli random variable if it has just two possible values. These two values are usually labeled numerically as 1 and 0, or in words as "success" and "failure", respectively.

The situation where this usually comes up is the following: We conduct **one** trial of an experiment and record the result as a success or failure. The related random variable X is given the value 1 if it was a success, and 0 if it was a failure. The words "success" and "failure" should not be taken too literally. They are just generic words used to describe the two possible results. For example, if we are studying a certain genetic disease and are observing one individual to determine if they have it, we might consider it a "success" if they do have it, simply because that was what we were looking for, not because it is a good thing to have the disease.

The distribution of a Bernoulli random variable is called a Bernoulli distribution. In the example of the genetic disease, let's say it is known that 2.6% of the population has the disease. Then when we randomly select one person and observe their disease status. We define X to be 1 if they have it and 0 if they don't. Then the pmf of X is:

$$\mathrm{P}(X = x) = \begin{cases} 1 - .026 & x = 0 \\ .026 & x = 1 \end{cases}$$

If we didn't happen to know that 2.6% number, things get slightly more interesting. In that case, we might use the notation p to represent the currently unknown proportion of people who have the disease. Then we would write the pmf of X as:

$$\mathrm{P}(X = x) = \begin{cases} 1 - p & x = 0 \\ p & x = 1 \end{cases}$$

This is our first example of a situation where there is a parameter (namely p) in our probability distribution. A parameter is a value that helps us define the distribution correctly, but that we don't know. In real situations, our goal is often to estimate this parameter. Here, an estimate of the parameter p would give us a guess at what proportion of the population had the disease, something which was previously unknown to us.

2.6.2 Binomial Random Variable

With regard to the previous section, it would be mostly uninteresting if all we did was sample a single individual, or more generally, conduct any experiment just one time. So, in real situations, a Bernoulli random variable is really just a means to a more interesting end. The "end" in this case is called a Binomial random variable.

In a real situation, we often conduct experiments a number of times. We call each repetition a *trial* and let n represent the total number of trials. A trial is a general word to represent many possible situations: a trial may be a randomly selected individual, or a single nucleotide, a single meiosis, a single mating, etc.

Suppose we do conduct an experiment for n trials (where we determine n in advance), and each trial is recorded simply as a success or failure as with a Bernoulli random variable. Also, for this discussion, it also needs to be true that:

1. Each trial is independent (i.e., the result of one trial will not affect the result of any other).
2. The probability of getting a success from one trial to the next is exactly the same. We will call this common success probability p.

A random variable that is often of interest to us in such a situation is the count of how many successes we had in those n trials. Let X be this random variable. Then X is said to be a Binomial random variable, and its probability distribution is called the Binomial distribution. You should notice that a Binomial random variable really just comes about from taking n Bernoulli random variables and adding them up.

Notice that the set of possible values for X is $\{0, \ldots, n\}$. If we were to want to write down the pmf of X, it could take up a lot of paper if n were even moderately large. Fortunately, the pmf of a Binomial random variable (and all the other we will discuss) can be written very nicely using a formula. Skipping over the details of where the formula comes from, the pmf of a Binomial r.v. is:

$$\mathrm{P}(X = x) = \binom{n}{x} p^x \ (1-p)^{n-x}, \quad x = 0, \ldots, n \tag{2.4}$$

Now, instead of a listing of values and probabilities, we can use this formula to calculate the probability of any particular value of X occurring (i.e., observing any particular number of successes) when I conduct these n trials.

Example 2.2: In our disease example, let's say again that we know the proportion in the population is 2.6%. Then let's randomly select 10 people who are independent of one another. For each of the 10 I will observe whether or not they have the disease. X= the number of the 10 that have the disease is a Binomial random variable. In this example, the p in the formula is .026 and the n is 10. So for this specific X, its pmf is:

$$\mathrm{P}(X = x) = \binom{10}{x} (.026)^x \ (1 - .026)^{10-x}, \quad x = 0, \ldots, 10$$

The chance that we will find that two of the ten have the disease is $\mathrm{P}(X = 2) = \binom{10}{2} (.026)^2 (1 - .026)^8 = .0246$. That's a low probability, but that's what we might expect given the problem setup.

■

As a reminder, once again it is more typical of real problems that we don't know a value like p ahead of time. Our objective would be to model the situation using Equation (2.4), run the experiment and collect some data, then estimate p based on evidence we find in our data. More on this later.

We often use a shortcut notation to describe the special distribution that a random variable has. For example, in the Binomial case, we would have to say "the random variable X has a Binomial distribution with n trials and p probability of success on each trial" in order to convey all the information we need about X. However, that is a lot to say, and so a simpler way to write the same thing is:

$$X \sim \mathrm{Bin}(n, p).$$

We'll see similar shortcuts for other random variables. The general setup of this notation is to write the random variable (here X), the "$\sim$" sign which means "has the following distribution", shorthand for the name of the distribution (here "Bin" for Binomial), and then the parameters of that distribution (here n and p). The parameters are a very necessary and important part of this notation because without them, we wouldn't know what to use in the Binomial pmf formula.

For our disease example, we would write: $X \sim \mathrm{Bin}(10, .026)$. It is generally assumed that one will understand the order in which the parameters are written. In other words, no one will confuse the situation and think you meant that $n = .026$ and $p = 10$ in this case. For other random variable types, we might need to be a little more careful about the ordering.

Since a variable which is Binomial is just a special kind of discrete random variable, we can use Formula (2.1) to calculate its mean and Formula (2.3) to calculate its expected value. The math to do this is a little tricky and we don't need to focus on that. The result is: if we have $X \sim \mathrm{Bin}(n, p)$, then

$$\mu_X = np \qquad \text{and} \qquad \sigma_X = \sqrt{np(1-p)}.$$

For the disease example, we would calculate $\mu_X = 10 \times .026 = 0.26$. Our interpretation would be that if we continually sampled 10 independent people, and for each sample of 10 recorded the number who had the disease, the average of all these recordings would be 0.26.

2.6.3 Geometric Random Variable

An interesting way to switch around the way we conducted the experiment described above is the following. Instead of deciding ahead of time how many trials we will conduct, we will keep conducting the experiment

(keep conducting trials), until we finally get one success. Other than that, the rest of our assumptions about the experiment are the same (i.e., the trials are independent and the chance of success is the same for each trial).

In this situation, you should notice that $X =$ Number of Successes over all the trials is not a random variable any more. We know its value is exactly one, because that's how we defined the experiment. But, what we don't know is how many trials we will wind up needing to conduct. So, in this case, we consider the number of trials as the random variable. We may get our one success on the first trial, in which case n would be 1. Or we may have 5 failures before finally getting the first success. Here n would be measured as 6. Using our usual notation for random variables, we will use N to represent this random variable, and n to represent its possible values.

With this setup, the random variable N is called a Geometric random variable. It is of course discrete since its possible values are integers. To be more specific, the set of possible values for N is $\{1, 2, \ldots\}$. Notice that N can't possibly less than one, because we have to attempt at least one trial to get a success. Also, we can't really put an upper limit on how large N can get, because we don't know when that first success might occur. In most situations, it might be extremely unlikely that N will be in the hundreds or thousands, but we can't rule it out probabilistically. Practically, there might also be physical limitations on how many trials could be conducted as you are waiting for the first success, as will be true in the example below. But, still, we'll consider that there is no upper limit on the possible values for N.

The notation we'll use to represent a Geometric r.v. is $N \sim$ Geometric(p), where p still represents the common success likelihood on each trial.

To calculate probabilities of the different values the could occur, we need to know the probability distribution of N. It is called the Geometric distribution, and the formula is:

$$\mathrm{P}(N = n) = p\ (1-p)^{n-1}, \quad n = 1, 2, \ldots \tag{2.5}$$

Example 2.3: Suppose we are analyzing the DNA sequences of a selection of genes in some species. We might be interested in the underlying chemistry of these genes and want to study the particular sequences of nucleotides that tend to occur. Specifically, let's say that whenever there is a T nucleotide in the gene, we would like to develop a model regarding the number of nucleotides until the next T. Consider the following part of a DNA sequence:

...AAGTGGGAGGCCATCCA...

Once we encounter the first T (the one in the fourth position), how long will it be until we encounter the next T? Notice we are defining a success as there being a T in a particular position. This is a random process, and that count of nucleotides is a random variable. It will result in a different value depending on which

T is the starting point, which gene you have sequenced, and other factors. For this particular sequence, we would observe $N = 10$ since our first success (another T) occurred at the 10th position from the starting point. The number of nucleotides until the next T after that is encountered could then be measured and considered another observation of N.

So, let N be the random variable we are measuring. Does N as described in the example have a geometric distribution? We can consider each trial to be the process of observing what the next nucleotide is. We might wonder, then, if each trial is then independent of the next, and if the chance of getting a success (i.e., noticing a T) is the same from trial to trial (nucleotide to nucleotide). In order to claim that this random variable N has a geometric distribution, we would have to be willing, at least for now, to believe those assumptions. Let's do that for this discussion.

What value of p should we use? For now, we'll say $p = .25$. This makes some sense, since at first glance we might think that since T is just one of four possible nucleotides, it should have a 25% chance of occurring in any particular position. In fact, maybe our objective in this study is to start out claiming $p = .25$, and then collecting data to see if that is a good reflection of reality.
So we now have $N \sim$ Geometric(.25). Probabilities for different values of N can be calculated using Formula (2.5):

$$\mathrm{P}(N = n) = .25\ (.75)^{n-1}, \quad n = 1, 2, \ldots.$$

For example, the probability that the next T will occur at the next position, namely two T's in a row is $\mathrm{P}(N = 1) = .25 \times .75^0 = .25$. The probability that the next T will occur in the second position after the start is $\mathrm{P}(N = 2) = .25 \times .75^1 = .1875$, and so on. These suggest, for example, that if our model of this situation is correct, about 25% of the times the next T should occur immediately after the last, about 18.75% of the times the next T should occur two positions later, and so on.

Now, if we set out and collect some data, that is, observe various values of N, we may be able to validate, or invalidate, our probability model. For example, if in 20 random observations of N, we notice N=1 18 times, then we might have some reason to believe that our model was not correct. We would have expected $N = 1$ to occur about 5 times. Not necessarily exactly 5, but certainly not 18.

■

To conclude the discussion of Geometric random variables, we can calculate the mean and standard deviation using the following formulas:

$$\mu_N = 1/p \qquad \text{and} \qquad \sigma_N = \sqrt{\frac{1-p}{p^2}}.$$

2.6.4 Negative Binomial Random Variable

Another common special kind of discrete random variable is called the Negative Binomial. It is actually just a generalization of the Geometric random variable (or you can think of the Geometric as a special case of the Negative Binomial). Compared to the Geometric, the only difference is that instead of counting the number of trials it takes to finally get our first success, we will count the number of trials it takes to get our r^{th} success. That is, we'll specify ahead of time a number r that represents how many successes we want to wait for, and then continue conducting trials until we get that many.

The random variable we care about is still N, the number of trials it takes to get those r successes. We will write $N \sim \text{NegBin}(r, p)$. Its probability distribution is given by:

$$\text{P}(N = n) = \binom{n-1}{r-1} p^r \ (1-p)^{n-r}, \quad n = r, r+1, \ldots.$$

Notice that the possible values of N in this case begin at the integer r. It can't possibly take fewer than r trials to get the r^{th} success.

Example 2.4: An example of a Negative Binomial distribution would be to consider the process of comparing two DNA sequences. Given a starting point, we might be interested in comparing the sequences base by base until there have been, say, 4 mismatches. The number of nucleotides we encounter along the way would be a random variable with a Negative Binomial distribution where $r = 4$. Notice that here we are using the word success to mean encountering a mismatch. The value of p in the distribution would be a reflection of how dissimilar the two sequences are. A larger value of p would reflect a model where we believed the sequences are less similar, because we would encounter successes (i.e., mismatches) more often. In this case, we would be more likely to observe smaller values of N, because we are more likely to come across that 4th mismatch sooner. Reverse the argument for smaller values of p. In practice, we might leave p as an unknown parameter in this distribution and try to estimate it based on data we collect, or we might specify a value for p and try to determine if this reflects reality correctly or not.

■

For a Negative Binomial random variable, we calculate the mean and standard deviation as:

$$\mu_N = r/p \qquad \text{and} \qquad \sigma_N = \sqrt{\frac{r(1-p)}{p^2}}.$$

2.6.5 Poisson Random Variable

Yet another common and important type of discrete random variable is the Poisson random variable. Such a random variable is actually quite related to a Binomial random variable, but we won't go into those details. For our sake, let's consider situations where we count up the number of times an event happens over some continuous interval, such an interval of time, or an interval of space. Also, it must be the case that the following two conditions are true:

1. Whether or not the event occurs in a particular subinterval is independent of whether or not the event occurs in a different subinterval.
2. The chance of the event occurring in a certain subinterval only depends on the length of that subinterval, not on where that subinterval occurs within the overall interval.

In this situation, this count is a discrete random variable, and is said to have a Poisson distribution. There is also a parameter λ that goes along with a Poisson distribution. This specifies the average number of times that we would expect that event to occur in that interval.

Our notation, then, for a Poisson random variables is $X \sim \text{Poisson}(\lambda)$, and its probability distribution is given by:

$$\text{P}(X = x) = \frac{e^{-\lambda}\,\lambda^x}{x!}, \quad x = 0, 1, \ldots. \tag{2.6}$$

The possible values of a Poisson random variable X are all the non-negative integers. In other words, it could be that we don't observe any such events in the interval (in which case X would be 0), or maybe we observe one event, or any other positive number of events.

Example 2.5: In a previous example, we considered a random variable X which was the count of the number of crossovers that occurred during a particular meiosis (between two loci A and B on the same chromosome). At the time, we modeled the probabilities of particular values of X using a generic discrete distribution. Now, with our knowledge of Poisson distributions, we might consider X as being a Poisson random variable since it is a count of the number of events (i.e., crossovers) that happen in a certain interval of space (i.e., the length of the chromosome between A and B). We would need to be sure that the two conditions listed above are true, or at least close to true. In this case, we might consider condition two as certainly true, and condition one as close enough to true to continue using the Poisson model.

So, now we would model probabilities for X using Equation (2.6). We would likely leave λ as an unknown parameter, signifying that we don't yet know the average number of crossover events that occur between A and B. That is something we will be trying to estimate as we attempt to gather a better understanding of the genetics underlying these two linked loci.

■

For a Poisson random variable, the mean and standard deviation are:

$$\mu_X = \lambda \qquad \text{and} \qquad \sigma_X = \sqrt{\lambda}.$$

Based on our definition of λ above, it should be no surprise that this is the mean of the random variable. Interestingly, it turns out that the standard deviation is the square root of the mean for a Poisson random variable.

2.6.6 Multinomial Distribution

In one way, the multinomial distribution is very similar to the Binomial distribution, but very different in another. The similarity is with respect to the situations where it will apply. We will use the multinomial distribution when we are conducting multiple trials of an experiment, where each of the trials are independent of one another. The key difference is that, for the multinomial situation, each trial can have any number of possible outcomes, instead of just two. This is where the prefix "multi" comes in, as compared to the prefix "bi".

We can now setup the notation for our multinomial situation. Since the n trials can each have any number of possible outcomes, let's generally say that there are k such possible outcomes and label them $\mathrm{O}_1, \ldots, \mathrm{O}_k$. With the binomial distribution, $k = 2$ so we would have just had an O_1 and O_2. The probabilities associated with these outcomes are denoted $p_1, \ldots, p_k$. In other words, for example, the probability that outcome O_3 will be the one that occurs on a particular trial is p_3. By necessity, it must be true that $\sum_i p_i = 1$ since one and only one outcome will occur on any trial.

There are, in fact, k random variables that we consider in this case, not just one. We will refer to them as $X_1, \ldots, X_k$, and they represent the observed counts from n trials of each category $\mathrm{O}_1, \ldots, \mathrm{O}_k$, respectively. Similar to our discussion above about the associated probabilities, it must be that $X_1 + \cdots + X_k = n$. We say that, together, the random variables $X_1, \ldots, X_k$ have a Multinomial distribution on n trials with parameters $p_1, \ldots, p_k$. Or we may write

$$(X_1, \ldots, X_k) \sim \text{Multinomial } (n, p_1, \ldots, p_k).$$

As another analogy to the Binomial distribution, recall that we used p to represent the probability of a "success" (i.e., the first possible outcome, or O_1) and $1 - p$ to represent the probability of a "failure" (i.e., the second possible outcome, or O_2). We could have labeled them p_1 and p_2 at the time, but since they must add up to 1, the p and $1 - p$ notation makes this more apparent. For the multinomial, the probability associated with the last of the outcomes is often written as $p_k = 1 - p_1 - \cdots - p_{k-1}$ for the same reason.

The probability distribution for multinomial random variables is often called more specifically a *joint* probability distribution since it involves more than one random variable. The formula is written as:

$$\mathrm{P}(X_1 = x_1, \cdots, X_k = x_k) = \frac{n!}{x_1! \cdots x_k!} \, p_1^{x_1} \cdots p_k^{x_k} \tag{2.7}$$

Example 2.6: Suppose we are collecting genotypic data for a gene which has two alleles, A and a. Consider the offspring of a mating between two heterozygotes at this gene. As we know, there are three possible genotypes for the offspring: AA, Aa, and aa. We have also seen that we would expect these to occur with probabilities 1/4, 1/2, and 1/4, respectively. Of course, for any fixed number of offspring (say, n) from this mating, the observed counts won't follow those probabilities exactly.

In fact, the observed counts in this case are random variables. We might refer to the number of AA's

we observe as X_1, the number of Aa's as X_2, and the number of aa's as X_3. Since the offspring of this mating are independent of one another with respect to their genotype (assuming no identical twins), we can say that $(X_1, X_2, X_3) \sim$ Multinomial $(n, \frac{1}{4}, \frac{1}{2}, \frac{1}{4})$.

■

Since there are actually k random variables in this setup, there are k means and standard deviations that can be computed - one for each. The formulas are very similar to the binomial mean and standard deviation formulas:

$$\mu_{X_i} = np_i \qquad \text{and} \qquad \sigma_{X_i} = \sqrt{np_i(1-p_i)}$$

for $i = 1, \ldots, k$.

In our example, let's say that $n = 5$. We would calculate

$$\mu_{X_1} = 5 \times \frac{1}{4} = 1.25$$
$$\mu_{X_2} = 5 \times \frac{1}{2} = 2.5$$
$$\mu_{X_3} = 5 \times \frac{1}{4} = 1.25$$

2.6.7 Summary

Table 2.1 lists the distributions discussed in this section, along with a summary of the situations in which they are used.

2.7 Common Continuous Distributions

The distribution, or pdf, we decide to use for a random variable is an important decision, as we've seen. It defines everything about the random variable. Like with discrete random variables, there are some common distributions that are used very often in real situations to model probabilities. We will discuss a few of them here. There are also some special continuous distributions that come up quite often in a different way. They aren't necessarily used to model probabilities of real-world phenomena, but are used as derived distributions that we need in the process of doing statistical inferences (such as hypothesis tests). We will differentiate between these two cases below. But, for either type, they are distributions for continuous random variables and so follow all the same notation and requirements that we have discussed.

First, we will discuss a few continuous distributions that often serve as probability models for real-world random variables. Like with discrete distributions, these probability functions are defined by parameters which

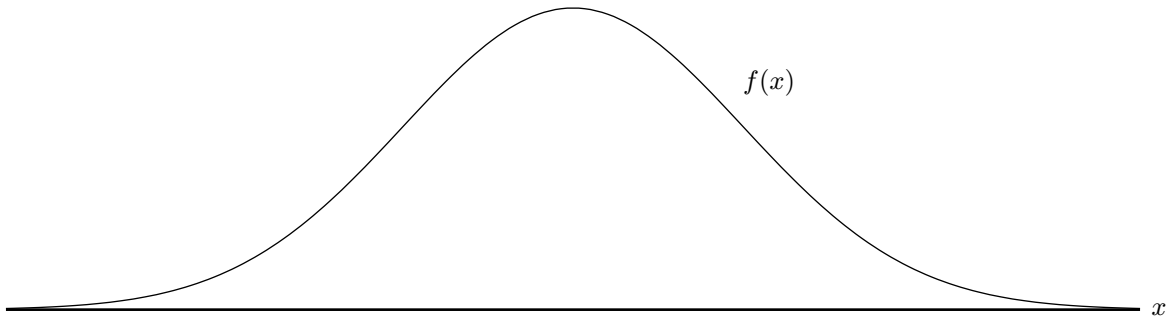
Table 2.1: Summary of common discrete probability distributions

Bernoulli	• One trial conducted - either a success or failure. • X=0 if failure; X=1 if success.
Binomial	• n trials • Each trial either a success or failure. • Trials are independent. • Probability of success p is same for each trial. • X=Total number of successes
Geometric	• Conduct trials until first success • Each trial either a success or failure. • Trials are independent. • Probability of success p is same for each trial. • N=Total number of trials
Negative Binomial	• Conduct trials until r^{th} success • Each trial either a success or failure. • Trials are independent. • Probability of success p is same for each trial. • N=Total number of trials
Poisson	• Count number of events that occur over time or space • Subintervals are independent. • Probability of event occuring in subinterval depends only on length of subinterval. • X=Total number of occurrences.
Multinomial	• n trials • Each trial can have one of k outcomes. • Trials are independent. • Probability of each outcome is same from trial to trial • X_i=Number of times outcome i occurs ($i = 1, \ldots, k$).

are typically unknown in a real application. So, as before, typical problems we'll come across is the need to estimate these parameters from data that we collect, or to validate the probability model that we are using.

2.7.1 Normal Distribution

The *normal distribution* is typically used to model probabilities for a continuous random variable when we believe the pdf curve should be symmetric and bell-shaped. In other words, if the curve should look approximately like:



The height example from above was actually an example of a random variable with a normal distribution, as you can now see by looking back at the way we had drawn that curve. The x-axis is not scaled in the picture above because the exact look and placement of the curve depends on two parameters of the normal distribution which we will discuss below.

Mathematically, the pdf of a random variable with a normal distribution is

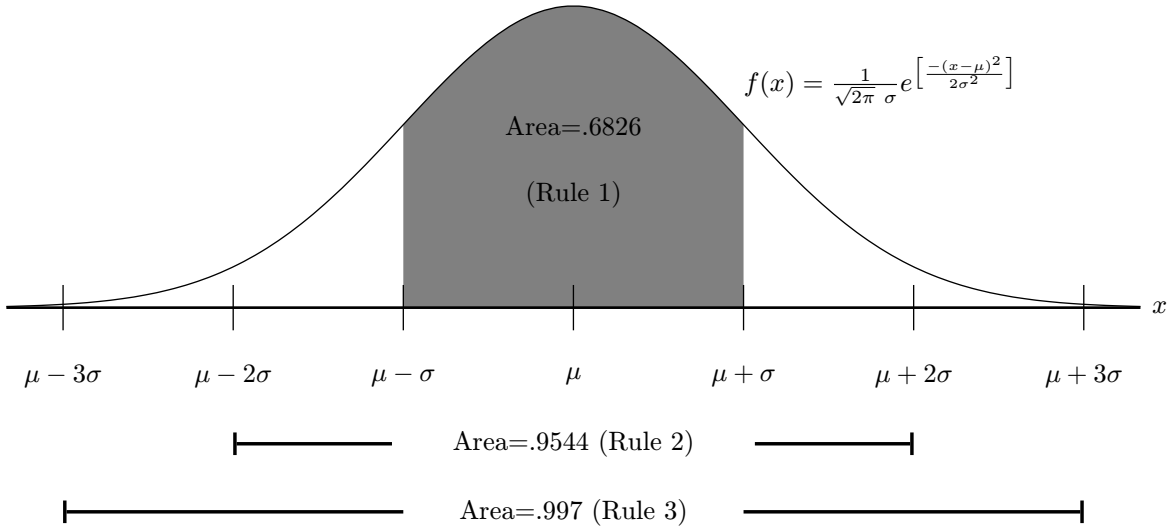
$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]}, \quad -\infty < x < \infty.$$

If a random variable X has this pdf, we will say that it has a normal distribution with parameters μ and σ, or write $X \sim \mathrm{N}(\mu, \sigma)$.

The parameters that completely define this distribution are μ and σ as mentioned above. If we analyzed the mathematical form of the pdf for X given above, we would notice that the curve will be centered on, and symmetric around, μ. So, μ defines where the curve is centered when we go and plot the pdf. A little more difficult to see is that the σ in the pdf defines how spread out the curve is. The larger σ is, the more spread; the smaller σ is, the less spread. More specifically, the following exact rules help define this amount of spread:

1. 68.26% of the distribution is within 1σ of μ.
2. 95.44% of the distribution is within 2σ of μ.
3. 99.7% of the distribution is within 3σ of μ.

With these rules in mind, we can better scale the horizontal axis on our plot of the pdf:



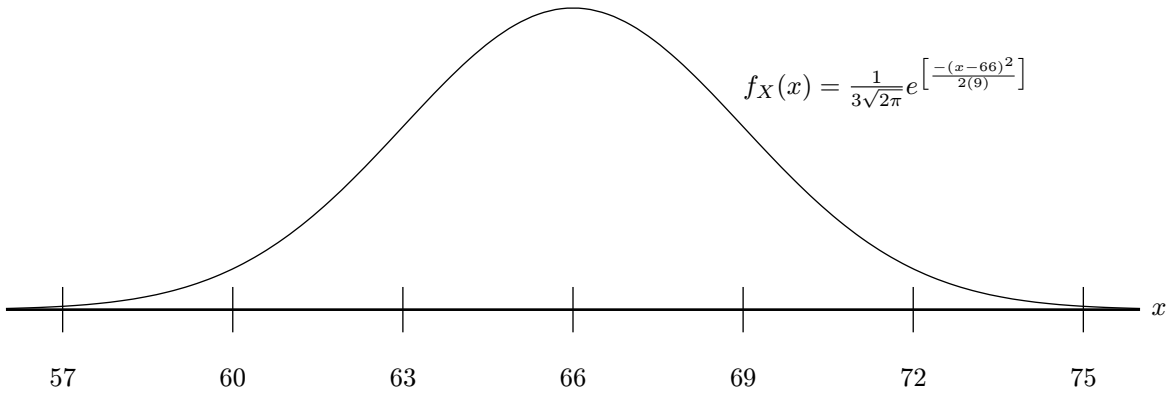
We typically use the third rule to help us accurately draw this curve for this function. Just draw it so that it has a bell-shaped symmetric look, is centered on μ, and is spread out in such a way that almost all of the area under the curve is between $\mu - 3\sigma$ and $\mu + 3\sigma$. Then mark off points on the x-axis appropriately.

Notice that the range of possible values for X is the entire real line. So, even though it might look like the curve $f(x)$ is coming down and hitting the x-axis on each side, it actually continues on forever and ever in each direction, getting closer and closer to the axis without ever touching it. Of course, on our small finite plot, we can't see this fact too well, but it important to realize it.

Example 2.7: Now, being more specific with our height example, we might define $X \sim \mathrm{N}(66, 3)$. So, the pdf of X is

$$f_X(x) = \frac{1}{3\sqrt{2\pi}} e^{\left[\frac{-(x-66)^2}{2(9)}\right]}, \quad -\infty < x < \infty.$$

and a plot of $f_X(x)$ looks like it did in our earlier discussion. Now, we have just put a functional form to the f we are using:



Some specific probability calculations we can make using our rules about the normal distribution are

$$\mathrm{P}(63 \leq X \leq 69) = .6826$$

$$\mathrm{P}(60 \leq X \leq 72) = .9544$$

and

$$\mathrm{P}(57 \leq X \leq 75) = .997$$

Of course, there are many other probabilities we might want to calculate that we can't answer using these rules, such as $\mathrm{P}(X >= 67.4)$ and $\mathrm{P}(58.6 \leq X \leq 62.93)$. These are still valid questions, but need to be answered by integration. In other words

$$\mathrm{P}(X >= 67.4) = \int_{67.4}^{\infty} \frac{1}{3\sqrt{2\pi}} e^{\left[\frac{-(x-66)^2}{2(9)}\right]} dx$$

This is not a fun integration problem to do, but thankfully we'll see later that we don't actually have to integrate in order to compute such integrals for normal distributions. Basically, all integrals of this type that we might be interested in have already been computed for us and tabled, as we'll see in Section 2.7.5.

■

Now, let's say a little more about the parameters μ and σ. It turns out that if we calculate the mean of a random variable with a normal distribution, we get by definition

$$\mu_X = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\ \sigma} e^{\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]} dx$$

and it turns out that this is equal to the parameter μ. So, this μ we have been referring to actually represents more specifically the mean of the random variable (which should be no surprise given that we were using the notation μ). Similarly, it turns out that the parameter σ represents what we might think; it is the standard

deviation of the random variable X.

Example 2.8: In the height example, we now know that $\mu_X = 66$ and $\sigma_X = 3$. Or, in words, the average height we would observe if we kept randomly selecting adults and measuring height forever and ever would be 66 inches. A measure of the spread of these heights would be 3.

Also, as a technical matter, notice that the normal distribution is actually not sensible from a pure mathematical standpoint to use as a probability model for height. This is because a normal probability model allows for the possibility that the random variable can have any real number value, including any negative number. Height, obviously, cannot have a negative value in the real world. However, the probability that would be assigned to negative values of height under this model would be so infinitely small that from a practical standpoint, we often use the model anyway.

■

The cdf of a normal random variable cannot be written down easily. We'll also deal with this more fully in Section 2.7.5.

2.7.2 Gamma Distribution

A random variable X is said to have a Gamma distribution with parameters α and β, or $X \sim \text{Gamma}(\alpha, \beta)$ if its pdf has the form

$$f_X(x) = \frac{1}{\Gamma(\alpha)} \frac{1}{\beta^\alpha} \; x^{\alpha-1} \; e^{(\frac{-x}{\beta})}, \quad x > 0; \alpha > 0; \beta > 0$$

and $f_X(x) = 0$ otherwise (i.e., for $x \leq 0$).
A couple of notes about this pdf. First, it only has positive probability for positive values of X. In other words, a random variable with a gamma distribution cannot have negative values. This is in contrast to the normal distribution, for example, where the random variable is allowed to have any real number value.

Second, we notice that the parameters α and β must also be positive real numbers. Otherwise, we would wind up with an ill-defined pdf (such as allowing for it to have negative values).

Third, the notation $\Gamma(\alpha)$ appears in this pdf. This is a mathematical function called the Gamma Function. It is defined as follows

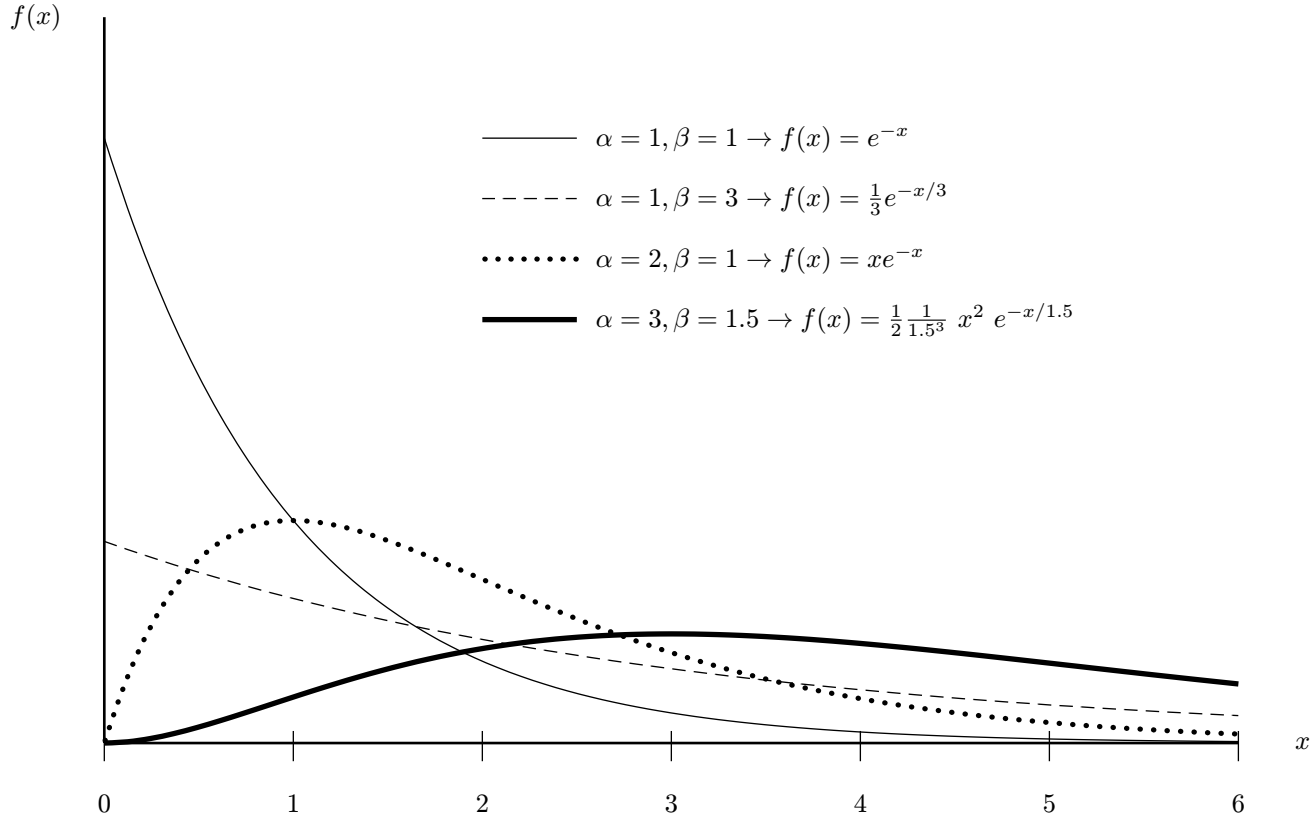
$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \; e^{-x} dx$$

The simplest way to think about this term in the Gamma pdf is that it is there to ensure that the function as a whole has area one under the curve. We do come across the need to calculate this function in some cases, and the following rules will often help us out

1. If $\alpha > 1$, $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$.
2. If α is an integer greater than or equal to 1, then $\Gamma(\alpha) = (\alpha - 1)!$.
3. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

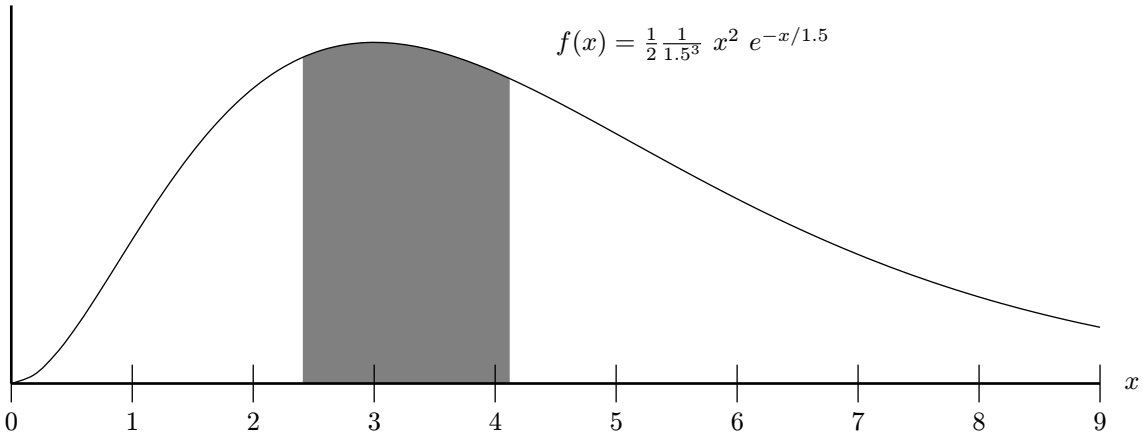
The first rule above establishes a recursive relationship for this function which is sometimes useful. The second defines that recursive relationship in a very simple manner (i.e., a factorial), if the α is an integer.

Now, if we plot this pdf, what shape does it have? That depends quite a bit on the particular values for α and β we use. Below are some examples:



The area under each of these curves is exactly 1, and each extends indefinitely to the right toward $+\infty$, getting closer and closer to the x-axis without ever touching it.From just this selection of four combinations of α and β, we can see that the gamma distribution can be used to model a wide variety of distribution shapes, and therefore a wide variety of real situations.

In some cases, calculating probabilities for gamma random variables is not too difficult with regard to the necessary calculus. For example, whenever $\alpha = 1$, the x^α term in the pdf drops out, and the pdf has the simple form of $\frac{1}{\beta}e^{-x/\beta}$ which is very quick to integrate over any range of x's. Other situations are not necessarily that immediate. For example, taking the case of $\alpha = 3$ and $\beta = 1.5$ above, we might want to know $\mathrm{P}(2.41 \leq X \leq 4.12)$:



This is of course defined as the integral

$$\mathrm{P}(2.41 \leq X \leq 4.12) = \int_{2.41}^{4.12} \frac{1}{2}\frac{1}{1.5^3}\ x^2\ e^{-x/1.5} dx$$

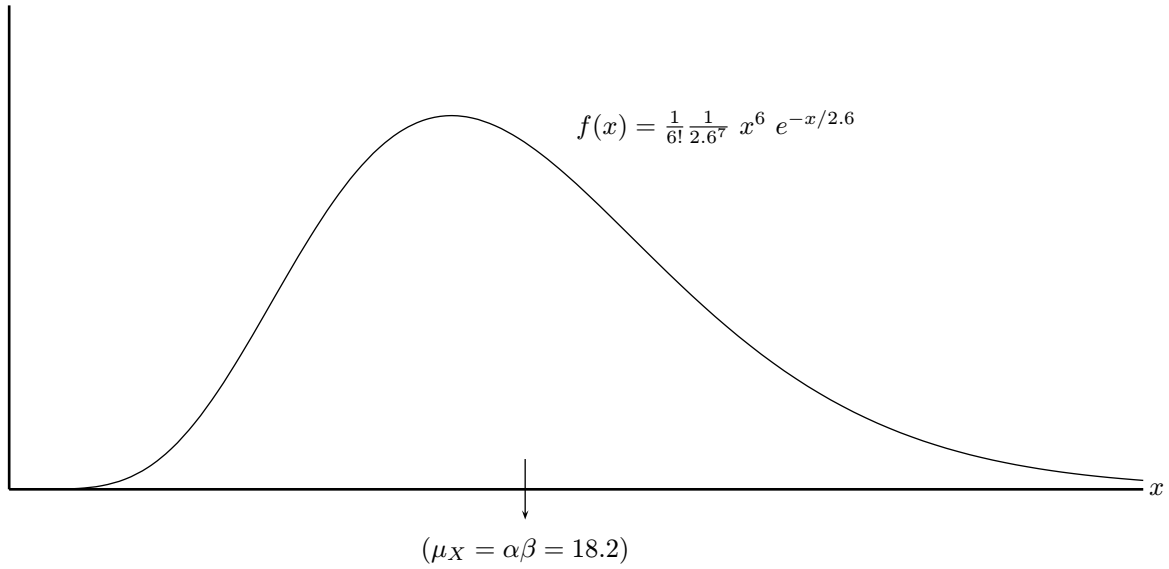
which would require integration by parts. If α were larger than 3, or it weren't an integer at all, this would turn into an ugly integral. Once again, we will not typically calculate these integrals by hand; we'll often rely on tables or computers to do this nasty work for us.

Let's say a little more about the effect of the parameters α and β on the shape of the distribution. It's not such a simple relationship as it was with μ and σ in the normal distribution. Both play a role in defining the mean and spread of the distribution, although β has a little more say in the spread than does α. Specifically, we can calculate the mean and variance of $X \sim \mathrm{Gamma}(\alpha, \beta)$ to be

$$\mu_X = \alpha\beta \quad \text{and} \quad \sigma_X^2 = \alpha\beta^2. \tag{2.8}$$

NOTE: These values are actually not difficult to calculate through integration. As an example, recall that $\mu_X = \int_0^\infty x \frac{1}{\Gamma(\alpha)}\frac{1}{\beta^\alpha}\ x^{\alpha-1}\ e^{(\frac{-x}{\beta})} dx$ by definition. Combining the x terms in the integrand and then setting up the integrand to look like a gamma pdf with different parameters will leave the constant $\alpha\beta$ outside of the integral, and the integral itself will integrate to 1. Try this on your own.

Finally, it is the case that a gamma random variable will have a distribution very similar to a normal distribution for larger and larger values of α and β. For example, if we have $X \sim \mathrm{Gamma}(7, 2.6)$, then the pdf of X looks like



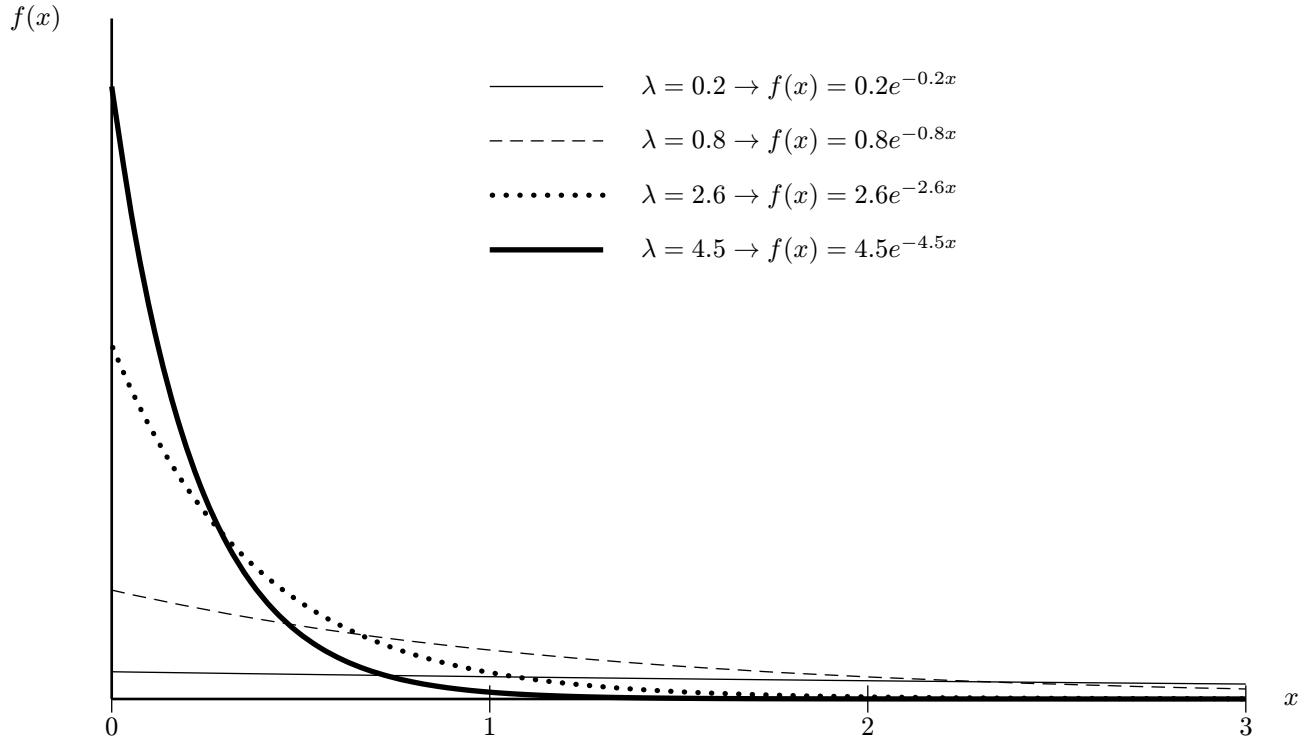
It is still somewhat skewed right, but is closer to symmetric than we saw in the earlier examples. Even larger values for α and β would be even more symmetric.

2.7.3 Exponential Distribution

The exponential distribution is actually a special case of the gamma distribution, and we have already seen some exponential pdfs. A random variable is said to have the exponential distribution with parameter λ, or $X \sim \text{Exponential}(\lambda)$, if its pdf has the form

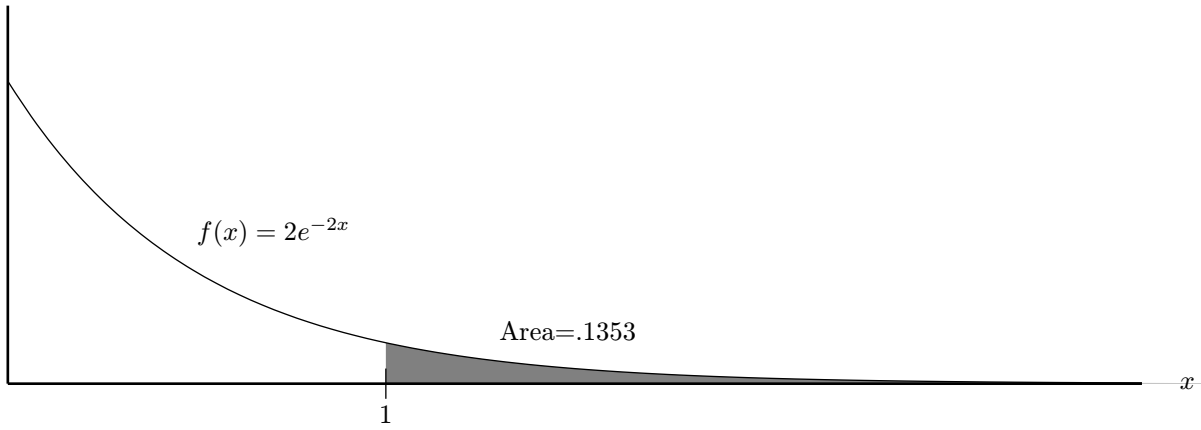
$$f_X(x) = \lambda e^{-\lambda x}, \quad x > 0; \lambda > 0$$

We can notice that this is in the form of a gamma pdf, where we have let $\alpha = 1$ and $\beta = 1/\lambda$ (since $\Gamma(1) = 1$ and the x term drops out). If you look back at the original pictures of various gamma pdfs, the ones with $\alpha = 1$ were actually more specifically exponential distributions. The pdfs of these distributions start off high on the x-axis and slope down in an exponential fashion. Some other examples are



The form of the exponential pdf is much simpler to deal with in general than the gamma pdf, so calculating probabilities is just a matter of simple calculus integrations. For example, say we have $X \sim$ Exponential(2.0). Then $f_X(x) = 2e^{-2x}$, $x > 0$. If we want to know the probability that X will have a value greater than 1, for instance, we would calculate

$$\mathrm{P}(X > 1) = \int_1^\infty 2e^{-2x}dx = -\big(e^{-2x}\Big|_{x=1}^{x=\infty} = -(0 - e^{-2}) = e^{-2} = .1353.$$



Because the exponential is just a special case of the gamma, formulas for the mean and variance follow simply from Equation (2.8):

$$\mu_X = 1/\lambda \quad \text{and} \quad \sigma_X^2 = 1/\lambda^2.$$

2.7.4 Beta Distribution

The beta distribution is a little different in nature than the distributions discussed so far. It is one in which the possible values allowed for the random variable are in the [0,1] range. A typical example of such a random variable would be a percentage or proportion, and an application in genetics would be the crossover percentage between a pair of loci. If we consider a randomly selected pair of loci on a chromosome, the crossover rate would be a random variable with range in [0,1].

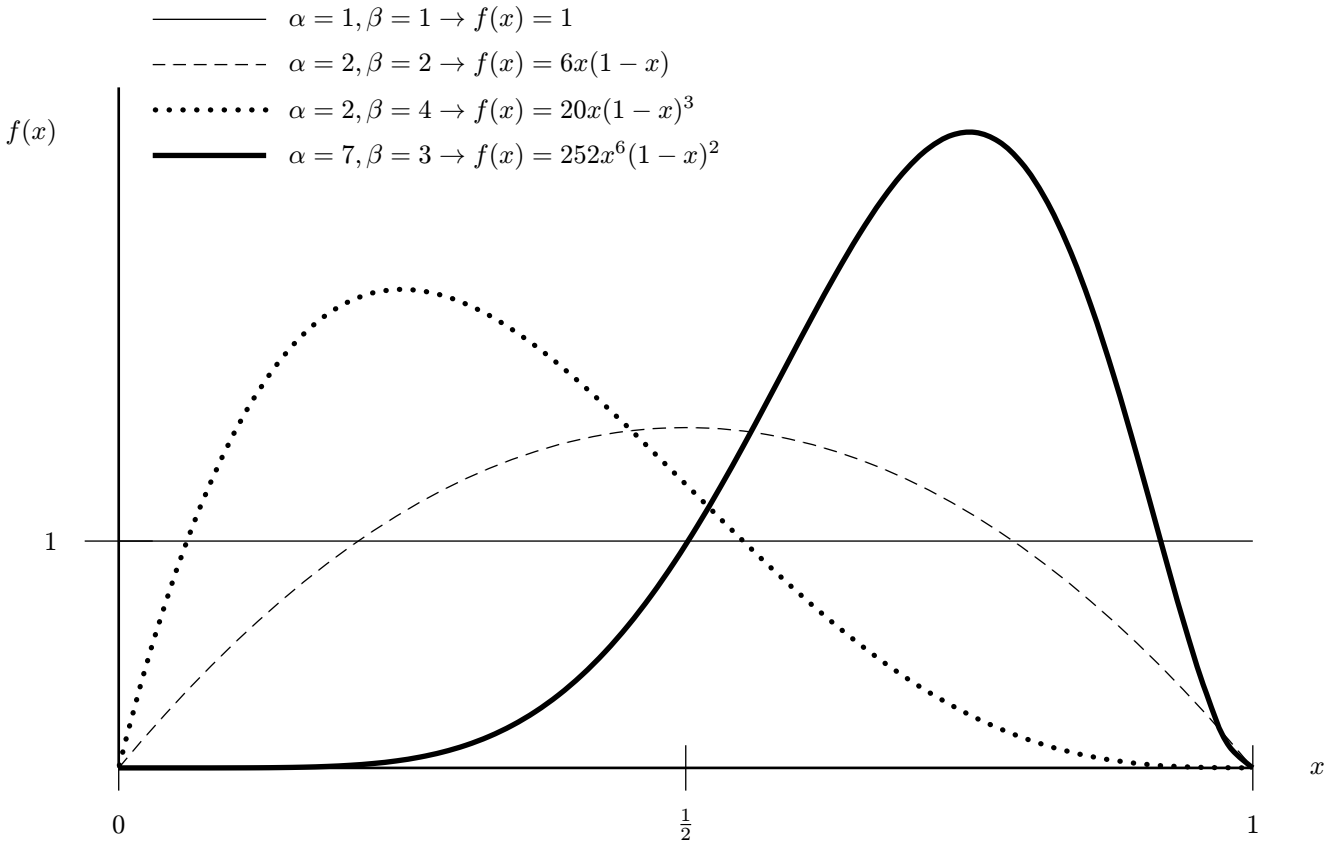
A random variable X is said to have the beta distribution with parameters α and β, or $X \sim \text{Beta}(\alpha, \beta)$, if its pdf has the form

$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}\, (1-x)^{\beta-1}, \quad 0 \le x \le 1; \alpha > 0; \beta > 0$$

where the Gamma function $\Gamma(\cdot)$ is as discussed earlier. The mean and variance of a beta random variable are

$$\mu_X = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \sigma_X^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Some examples of this pdf are below. Depending on the values used for α and β, this function gives a wide range of shapes in the [0,1] interval.



Probability calculations for the beta distribution can involve messy integrations, so we will rely on computers to do the work for us, as we'll see later in this chapter.

2.7.5 Standard Normal Distribution

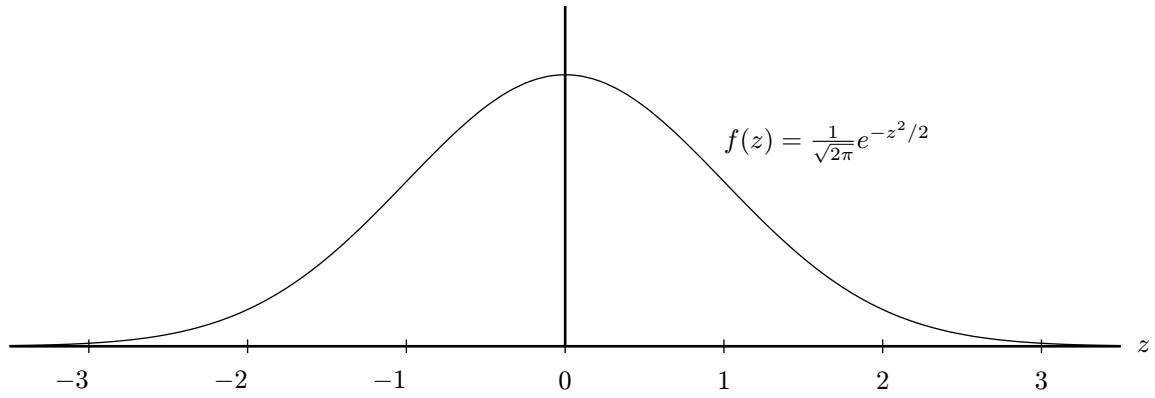
As mentioned earlier, there are a number of special continuous random variables that only come up in real applications as an intermediary variable derived from other random variables; they are not typically used as probability models for real data. They are used in many confidence interval and hypothesis testing situations, some of which we will come across later.

The most common of these distributions are the standard normal, t, chi-squared, and F distributions. We will discuss these below. Since they are not used as probability models, we will simply write their pdfs, plot the pdf for a couple of typical parameters, and make a couple of brief comments. We will mainly be concerned with finding percentiles of these distributions, which will be discussed briefly at the end of this chapter. More details on their use in applications will come up throughout the course.

Let's begin with the standard normal distribution. A random variable is said to have the standard normal distribution if it has a normal distribution with $\mu = 0$ and $\sigma = 1$. This distribution, then, is a (very) special case of the normal distribution. This type of random variable comes up so often that it is usually referred to with the letter Z, and the distribution is often called the Z-distribution. It's pdf is just the resulting special case of the normal pdf seen before:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty.$$

There are no parameters in this distribution, so there is one and only one standard normal distribution. The plot of this pdf is below. It is often called the standard normal curve, or the Z-curve.



By the rules seen earlier for a generic normal distribution, 99.7% of the area under this standard normal curve lies between the values -3 and 3. That is, $\mathrm{P}(-3 \leq Z \leq 3) = .997$.

One final note is a comment about where the standard normal distribution derives from. If we start with a generic normal random variable, say $X \sim \mathrm{N}(\mu, \sigma)$, and create a new random variable according to the transformation

$$\frac{X - \mu}{\sigma},$$

then this new random variable which is often referred to as Z will have a standard normal distribution. The process of creating this transformed random variable is often referred to as *standardizing* X.

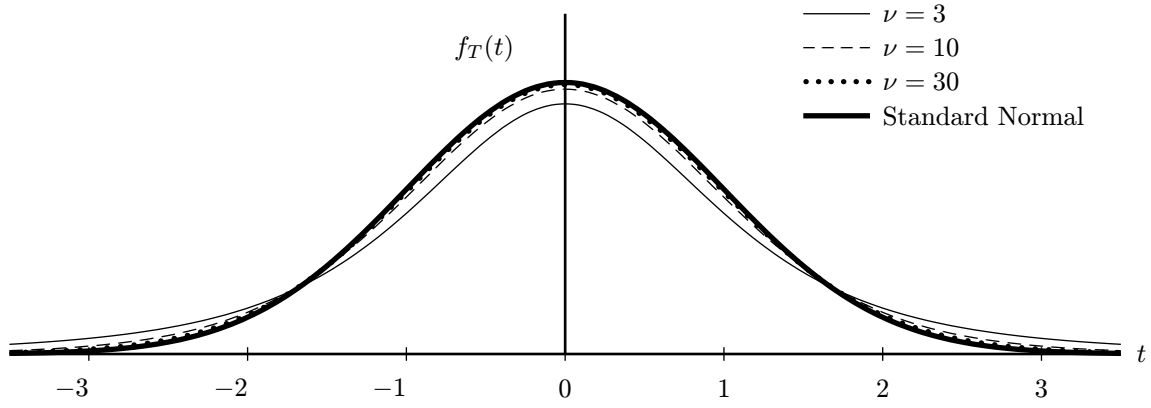
2.7.6 t-distribution

A random variable T is said to have the t-distribution with parameter ν, or $T \sim t_\nu$, if its pdf is of the form

$$f_T(t) = \frac{\Gamma[\frac{\nu+1}{2}]}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\big(1+\frac{t^2}{\nu}\big)^{-\frac{\nu+1}{2}}, \quad -\infty < t < \infty; \nu \text{ positive integer}$$

The parameter ν is often called a *degrees of freedom* parameter, or *df*, and is a kind of parameter typical of this and the other derived distributions we will discuss. It is a positive integer, and its value in a particular application depends on certain information about the data collection process; usually it is very closely related to the sample size. The t-distribution is also called Student's t-distribution, or the Student distribution.

Different values of the df lead to slightly different looking t-curves. As the degrees of freedom increases, the variance of the distribution, and therefore the spread of the curve, gets smaller. It is always symmetric about 0.



The plot above also includes a plot of the standard normal curve on the same scale (the darker solid line). We can see that the t-curves are actually quite similar to the standard normal curve. For smaller values of degrees of freedom, the curve has more spread than the standard normal, but as the df increases, the resulting t-curve begins to look more and more like the standard normal curve. Once the df gets to about 50, the two curves are the same for all practical purposes.

The mean and variance of a t-distribution are

$$\mu_T = 0 \quad \text{and} \quad \sigma_T^2 = \frac{\nu}{\nu-2}$$

again showing the similarity to the standard normal. Notice the variance goes to 1 as $\nu \to \infty$.

The t-distribution comes about in the following way: let $X \sim \mathrm{N}(0,\sigma)$ and $Y \sim \chi^2_\nu$ (discussed in the next subsection), and let them be independent. Then the random variable created by the following transformation

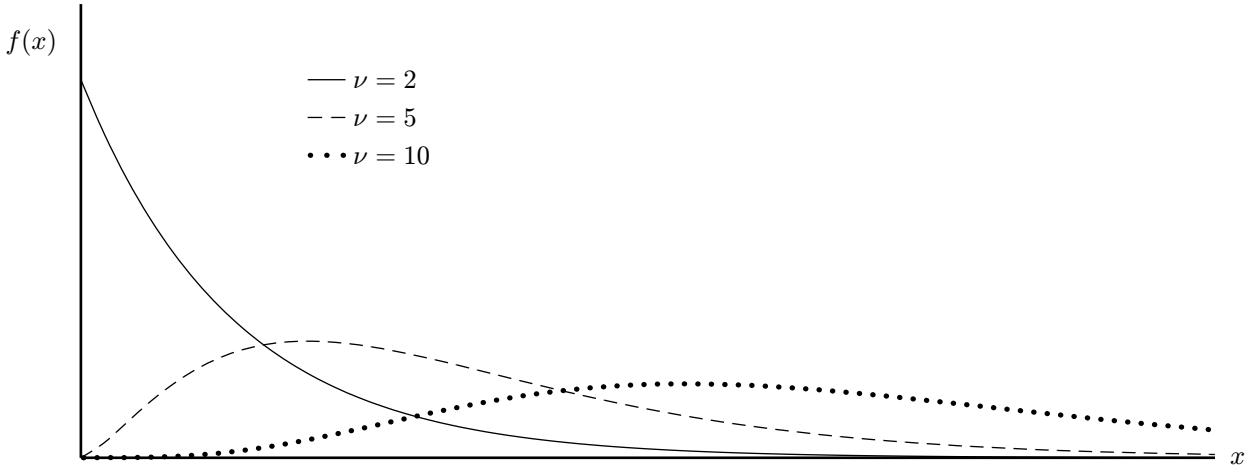
$\frac{X}{Y/\nu}$ has a t_ν distribution. It also arises in other related ways that we'll see.

2.7.7 Chi-squared distribution

A random variable X is said to have the chi-squared distribution with parameter ν, or $X \sim \chi^2_\nu$, if its pdf is of the form

$$f_X(x) = \frac{1}{\Gamma(\frac{\nu}{2})} \frac{1}{2^{\nu/2}} \, x^{(\nu/2-1)} \, e^{(-\frac{x}{2})}, \quad x > 0; \nu \text{ is a positive integer}$$

If you look closely, you'll notice that this is actually just a special case of the Gamma distribution, where we have let the α parameter be $\nu/2$ and the β parameter equal 2. As with the t-distribution, the ν parameter in a chi-square distribution is called its degrees of freedom and is a positive integer. Different values for ν lead to slightly different chi-square curves, with the main part of the curve shifting right as ν increases.



The mean and variance of a chi-squared random variable follow the formulas for the gamma distribution:

$$\mu_X = \frac{\nu}{2} \times 2 = \nu \quad \text{and} \quad \sigma^2_X = \frac{\nu}{2} \times 2^2 = 2\nu$$

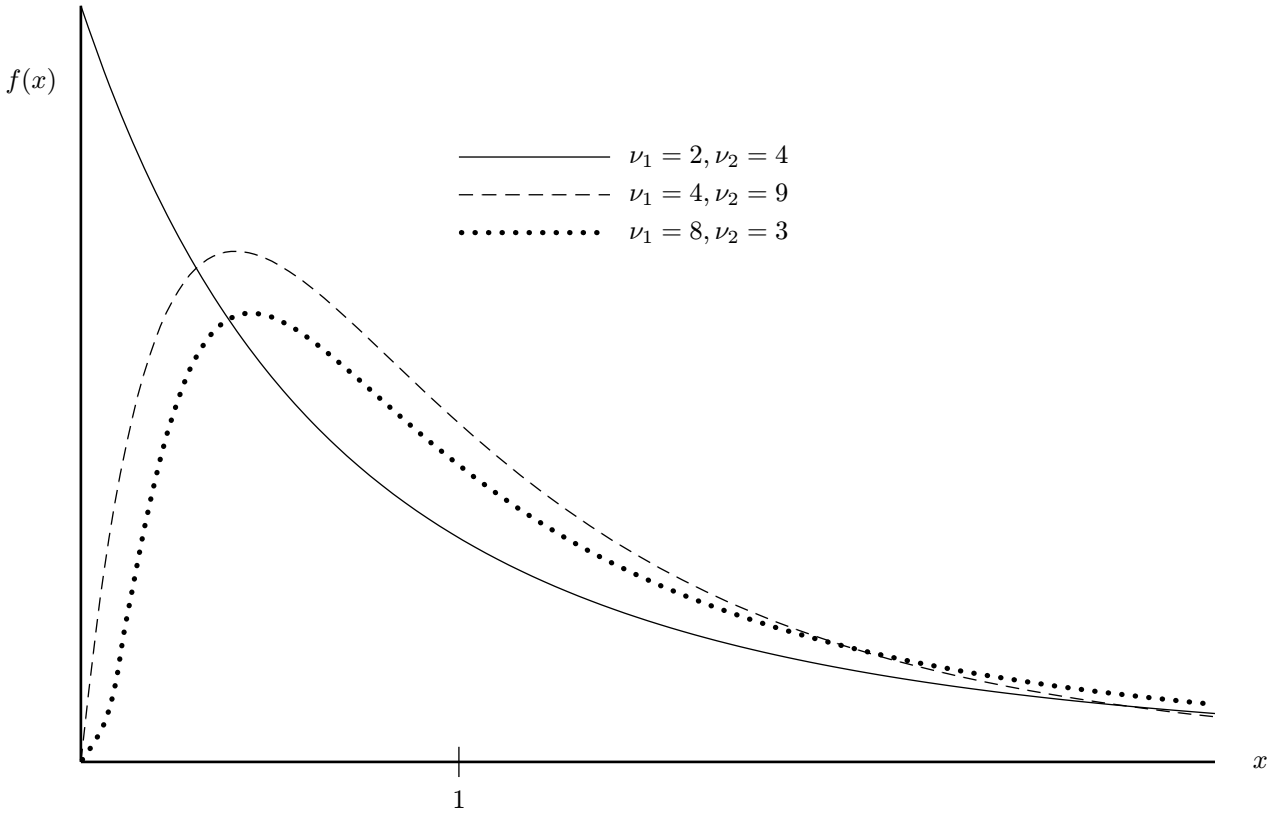
The chi-squared distribution arises in the following way. Let $Z_1, \ldots, Z_r \sim \mathrm{N}(0,1)$. Then the random variable created by summing the squares of these random variables, that is $\sum_i Z_i^2$, has a χ^2_r distribution.

2.7.8 F-distribution

A random variable X is said to have the F-distribution with parameters ν_1 and ν_2, or $X \sim \mathrm{F}_{\nu_1,\nu_2}$, if its pdf is of the form

$$f_X(x) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2}) \; \nu_1^{\nu_1/2} \; \nu_2^{\nu_2/2}}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \; \frac{x^{(\nu_1/2-1)}}{(\nu_2+\nu_1 x)^{[(\nu_1+\nu_2)/2]}}, \quad x > 0; \nu_1, \nu_2 \text{ are positive integers}$$

The parameters ν_1 and ν_2 are both degrees of freedom-type parameters. The order matters, so which is the first and which is the second is important. Often, the first is called the *numerator degrees of freedom* and the second is called the *denominator degrees of freedom.* Different values of these parameters lead to different looking F-curves:



The mean and variance are

$$\mu_X = \frac{\nu_2}{\nu_2 - 2} \quad \text{and} \quad \sigma_X^2 = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}$$

We can see that the mean tends toward 1 as $\nu_2 \to \infty$, and the variance tends toward 0 as both parameters increase.

The F-distribution comes about in the following way: let $X_1 \sim \chi^2_{\nu_1}$ and $X_2 \sim \chi^2_{\nu_2}$ and let them be independent. Then the random variable created by the transformation

$$\frac{X_1/\nu_1}{X_2/\nu_2}$$

as an F_{ν_1,ν_2} distribution.

2.7.9 Using the Computer for Distribution Calculations

Microsoft Excel is an easily accessible package that provides functions for computing probabilities and percentiles for many common distribution functions. Tables 2.2 and 2.3 gives a summary of these functions for various distributions, along with some examples and notation.

Also, the website http://calculators.stat.ucla.edu/cdf/ is a great resource for making such calculations if you are attached to the internet. From here, you can calculate cdfs, pdf/pmfs, random numbers, and also plot pdfs/pmfs.

Table 2.2: Excel Formulas for Discrete Distributions

Distribution	Probability (pmf) $P(X = x)$	Cumulative Probability (cdf) $F_X(x) = \mathrm{P}(X \leq x)$	Percentiles X_{pct}
Binomial(n, p)	BINOMDIST$(x, n, p$,FALSE)	BINOMDIST$(x, n, p$,TRUE)	CRITBINOM$(n, p$,pct$)-1$
NegBin(r, p)	NEGBINOMDIST$(x - r, r, p)$	–	–
Poisson(λ)	POISSON$(x, \lambda$,FALSE)	POISSON$(x, \lambda$,TRUE)	–

Table 2.3: Excel Formulas for Continuous Distributions

Distribution	Cumulative Probability (cdf) $F_X(x) = \mathrm{P}(X \leq x)$	Percentiles X_{pct}
Beta(α, β)	BETADIST(x, α, β)	BETAINV(pct,$\alpha, \beta)$
$\chi^2(\nu)$	CHIDIST(x, ν)	CHIINV(pct,ν)
Exponential(λ)	EXPONDIST$(x, \lambda$,true)	– (Use GAMMAINV)
F(ν_1, ν_2)	FDIST(x, ν_1, ν_2)	FINV(pct,$\nu_1, \nu_2)$
Gamma(α, β)	GAMMADIST$(x, \alpha, \beta$,TRUE)	GAMMAINV(pct,$\alpha, \beta)$
N(μ, σ)	NORMDIST(x, μ, σ)	NORMINV(pct,$\mu, \sigma)$
N(0,1)	NORMSDIST(x)	NORMSINV(pct)
t(ν)	1−TDIST$(x, \nu, 2)$	TINV$(1 - \mathrm{pct}, \nu)$

2.8 Multiple Random Variables

In most real applications, there is not just one random variable. In fact, there are usually many random variables in a problem. We have really already seen that through examples discussed so far, without usually explicitly saying so. For example:

1. A Binomial random variable is really just n Bernoulli random variables added up. In other words, if

$X_1, \ldots, X_n$ are each Bernoulli(p) random variables, then $X_1 + \cdots + X_n$ is a Binomial random variable as long as the independent trials assumption is true. The n Bernoulli's can be thought of as storing the result of each trial, either a 1 for success or 0 for failure. Adding them up gives the total number of successes, which is how we define a Binomial.

2. Recall the random variable $X \sim$ Poisson(λ) that was a count of the number of crossovers during a meiosis. It is unlikely that we'll observe just *one* meiosis in our research to better understand this genetic phenomena. We're likely to observe many such meioses. The resulting number of crossovers for each observation is a random variable with the same distribution. If we make n such observations, we have n random variables which we might refer to as $X_1, \ldots, X_n$, each having a Poisson(λ) distribution.

3. The multinomial distribution was a direct example of having multiple random variables. It is defined that way right from the start.

2.8.1 Random Samples and Independence

Situations 1 and 2 above are extremely common in statistical analysis and have special terminology associated with them. First, we say that the n random variables are *independent* of each other, or we might call them *mutually independent*. By saying two or more random variables are independent, we mean that the measurement we make for one will have no effect at all on measurements we make for any other.

The other commonality in Situations 1 and 2 above is that each of the n random variables is just another observation (i.e., measurement) of the same basic phenomena. They just represent different trials, or time periods, or meioses, etc. So, not only are they independent, but they each have the *same distribution*. For example, when observing n meioses and counting the number of crossovers each time between loci A and B, we are just observing different trials of the same basic phenomena. Therefore, if one of those random variables has a Poisson(λ) distribution, then they all must have that distribution.

Putting these two concepts together leads to the idea of a *random sample*. We say that n random variables are a random sample if they are mutually independent and each has the same distribution. "Having the same distribution" is often said as "identically distributed". The notation we will use is (replace "Poisson" with whatever distribution is being used in the problem)

$$X_1, \ldots, X_n \text{ are a random sample, each with a Poisson}(\lambda) \text{ distribution}$$

or, in shorthand:

$$X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$$

or:

$$X_1, \ldots, X_n \text{ iid Poisson}(\lambda)$$

where iid stands for independent and identically distributed.

For the time being, we should note that the random variables involved in a multinomial distribution are a good example of not being independent. Notice that knowing the value of one or more can very much affect the value of others, since there is the restriction that they must all add up to n. However, other than the multinomial, most situations we come across will fall into the category of independence.

2.8.2 Joint Probability Distributions

Before, with one random variable at a time, we just needed a probability distribution that described the likelihood of that one random variable's possible values. Now that we realize that we are often talking about many random variables at once, we need to be able to describe the probabilities of various combinations of their values occurring. Such a function is called a *joint probability distribution* for n random variables. More specifically, for discrete random variables, it is called the *joint probability mass function (joint pmf)*, and for continuous random variables, it is called the *joint probability density function (joint pdf)*.

For situations where we have a random sample, determining the joint pmf or pdf is, thankfully, quite simple. First, a simplifying notation that we'll use is the following. Instead of writing $\mathrm{P}(X = x)$ to represent the pmf of a discrete random variable, we write $\mathrm{p}_X(x)$. And we'll write the joint pmf of n random variables $X_1, \ldots, X_n$ as $\mathrm{p}_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$. For continuous random variables, we will write an individual pdf as $f_X(x)$ and a joint pdf as $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$.

Now, the result that we can make use of is the following: if $X_1, \ldots, X_n$ are a random sample, each with probability distribution $\mathrm{p}_X(x)$ or $f_X(x)$, then their joint pmf or pdf is just the product of the n individual distributions. In other words

$$\mathrm{p}_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathrm{p}_{X_1}(x_1)\ \mathrm{p}_{X_2}(x_2)\ \cdots\ \mathrm{p}_{X_n}(x_n)$$

for discrete variables, and

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1)\ f_{X_2}(x_2)\ \cdots\ f_{X_n}(x_n)$$

for continuous variables.

Example 2.9: Consider our meiosis example. We have $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$. We know the pmf of each of the individual random variables $X_1, \ldots, X_n$:

$$p_{X_1}(x_1) = \frac{e^{-\lambda}\ \lambda^{x_1}}{x_1!}$$
$$p_{X_2}(x_2) = \frac{e^{-\lambda}\ \lambda^{x_2}}{x_2!}$$
$$\cdots$$
$$p_{X_n}(x_n) = \frac{e^{-\lambda}\ \lambda^{x_n}}{x_n!}$$

So, the joint pmf of the random variables is:

$$\begin{aligned} \mathrm{p}_{X_1,\ldots,X_n}(x_1,\ldots,x_n) &= \frac{e^{-\lambda}\ \lambda^{x_1}}{x_1!} \ \cdots\ \frac{e^{-\lambda}\ \lambda^{x_n}}{x_n!} \\ &= \frac{e^{-n\lambda}\ \lambda^{(\sum_i x_i)}}{x_1! \cdots x_n!} \end{aligned}$$

■

Let's look ahead for a minute and ask, with respect to the above example, why all this will be important. In this situation, we are not claiming that we know λ. We have left it as an unknown parameter. Recall that the λ in a Poisson distribution represents the average number of times the event (here, crossovers) occur in the length of space or time (here, a length of DNA). We don't know λ, but it is a quantity of extreme importance if we are to better understand the genetics involving these two loci.

So, one ultimate goal of this study is likely to be that we want to come up with an estimate of λ. If we observe just one such meiosis, will this provide us the necessary data to make such an estimate? Would it be a reliable estimate? Definitely not.

On the other end of the spectrum, if we observed 100,000 such meioses, would this provide us with what we needed to make an accurate estimate of λ? It seems that the answer is certainly yes. A reasonable value for n needs to be chosen that provides a tradeoff between accuracy of the estimate we make, and the amount of work we do.

In summary, we need to make multiple observations of phenomena we are interested in for us to make any valid conclusions regarding that phenomena or to understand better how that phenomena really works.

Example 2.10: Let's also reconsider the geometric random variable example. In that case, we were counting the number of bases along a DNA strand, from some starting point, until and including the next T. This random variable X had a Geometric distribution with parameter $p = 1/4$.

Now, more realistically, we are likely to perform this counting process multiple times from multiple selected starting points. Each of these independent counting processes will result in the measurement of a random variable. Let's do this n times, and call the resulting random variables $X_1, \ldots, X_n$, as usual. Each has a geometric distribution with parameter $1/4$. Since they are a random sample, their joint pmf will be:

$$\mathrm{p}_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = (\frac{1}{4})(\frac{3}{4})^{x_1-1} \ \cdots \ (\frac{1}{4})(\frac{3}{4})^{x_n-1}$$

$$= (\frac{1}{4})^n(\frac{3}{4})^{\sum_i(x_i-1)}$$

■

We can use joint pmf's to calculate probabilities if we want. For example, in the geometric example above with $n = 4$, we may ask: What is the probability that our first trial will result in a count of 3 bases, then it will take 6 bases until then next T, then 2 bases, then 8 bases? This probability can be calculated from the joint pmf. It is just:

$$\mathrm{p}_{X_1,X_2,X_3,X_4}(3,6,2,8) = (\frac{1}{4})^4(\frac{3}{4})^{[(3-1)+(6-1)+(2-1)+(8-1)]} = .0000522$$

But, this is not our typical use of these pmfs, as we'll see.

2.8.3 Statistic

A *statistic* is a any function of random variables that does not involve an unknown parameter. A statistic (or statistics) is typically a way to summarize the result of a random sample with one (or a few) values. The most common statistic that we'll come across is one where we take the average of the random variables from the random sample. This statistic is written $\bar{X}$ and read as "X-bar". Mathematically, it is:

$$\bar{X} = \frac{\sum_i X_i}{n}$$

Notice that such a statistic gives us a quick summary of what happened in the overall experiment by providing us the average value that occurred. Other statistics that are often considered are

- $S = \sqrt{\frac{\sum_i X_i^2 - \frac{(\sum_i X_i)^2}{n}}{n-1}}$ the standard deviation of the values of the random variables. This is measure of how spread out the values were.
- The median of the random variables (the middle value).
- The smallest of the random variables.
- The largest of the random variables.

This is just a small list of statistics that may be used in any particular situation. They each serve a particular purpose depending on the situation. We will most often concentrate on $\bar{X}$, but we'll come across others as well.

The most important thing to recognize about a statistic is that it too is a random variable in its own right. In other words, the value we will get for $\bar{X}$ is not known to us ahead of time. It's value will depend on the number of observations, which observations we collected data on, the exact place and time observations were collected, etc. Since it is a random variable, it will have a probability distribution and we can discuss it just like we've discussed other random variables so far.

2.9 Analysis of Probability Models and Parameter Estimation

2.9.1 Overview of Problems to be Solved

For both discrete and continuous random variables, the general problem types we encounter are the same. The two classes of problems can be thought of as:

- Given a probability model, validate its correctness based on how well it describes real data.
- Given a probability model that we assume to be correct, estimate the unknown parameters of the model based on data that has been collected (or conduct tests regarding those parameters).

Here we will give a general overview of both of these for a continuous random variable example, and then discuss the second in more detail.

Let's set up the example. Let X be a continuous random variable that measures the amount of time it takes for a certain cellular protein to degrade. Protein degradation can occur spontaneously due to the protein's aging process; eventually, the molecular structure may degrade and no longer function. Or, a protein may be actively degraded by other molecules in the cell, such as other proteins (see the Ewens and Grant reference).

X is certainly a random variable since any particular observation of this process will yield different times due to random and unknown factors in the cellular process. As researchers in this area, we are interested in constructing a probability model that does a good job of describing degradation for this particular protein.

Let's start out considering the possible models. Since X is a time measurement, we are likely to consider probability models that take this into account (i.e., allow for the random variable to take on non-negative real numbers). Although there are many such probability models, a couple of common ones that we have discussed are the gamma distribution and the exponential distribution (which is just a special case of the gamma). Both are actually very common distributions to use to model waiting times (in this case, waiting time until the protein degrades).

For this example, let's say that we will use the exponential distribution. So, we formally define our model like this:

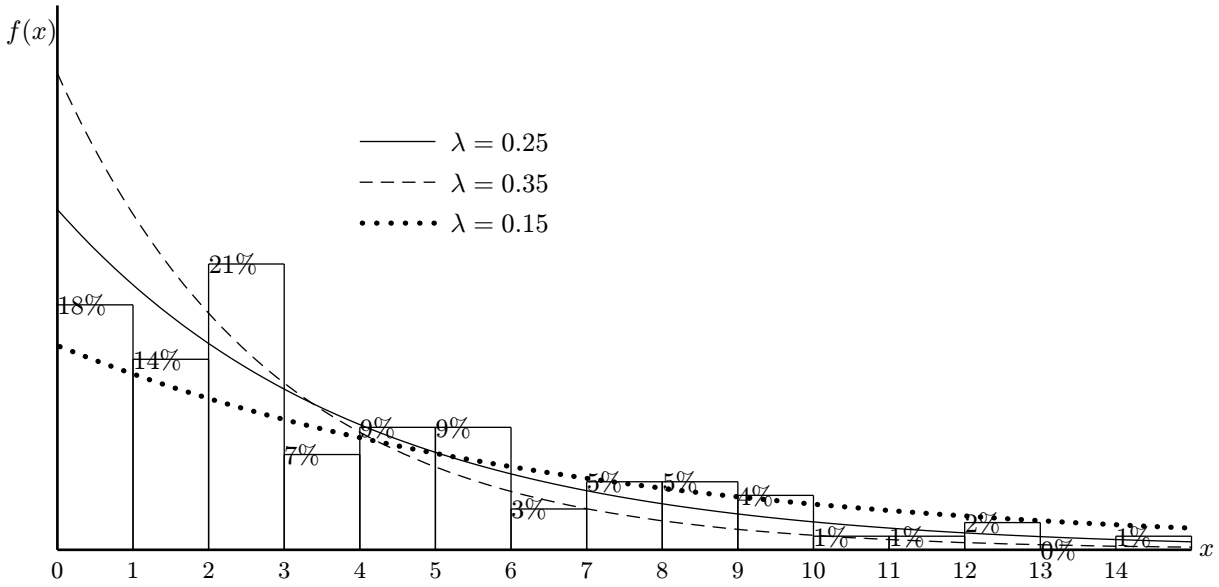
$$X_1, \ldots, X_n \sim \text{Exponential}(\lambda).$$

This states that we will take n independent observations of this random variable, and that measured values will follow an exponential distribution with unknown parameter λ. Instead of saying "an exponential distribution," we might more correctly say "*some* exponential distribution". In other words, we are claiming that the model follows the basic shape of an exponential distribution (some examples of which are in the relevant section earlier in this chapter), but since we don't know λ, we don't know for sure which such exponential distribution is correct. So, our first goal is to estimate λ based on data we collect.

Say we collect data on $n = 100$ randomly selected such proteins. Those 100 values are below:

```
 1.0488   8.7128   5.1088   2.4621   8.0735   4.4189   4.2664
 5.0919   5.5437   1.6770   1.7151   1.7572   0.5448   2.6872
 4.2220   1.0151   5.3031   3.8807   2.0359   0.0426   1.4084
 6.7384   0.1361   2.0712   3.3359   0.6761   2.7327   4.2882
 4.4953   1.8219   1.2427   8.9060  12.2685   2.8841   0.5807
 0.5927   3.4842   0.9928   9.4022   3.8196   6.3884   2.1940
 0.2338   2.1911   0.9576   3.9605   2.7559   6.6619   2.2530
 4.2422   0.7012   1.1716   5.0964   1.8256   0.4665   0.9726
 2.0984   7.5901   1.1320   7.7666   7.2670   9.8356   1.1392
 0.5069  11.2457   5.7494  10.2586   2.5182   1.9329   9.8432
 2.0279   0.6773   8.9189   2.3682   8.1927   7.2889   7.9581
12.2379   2.9023   2.3436   2.4534   0.7263   1.4295   0.8745
 2.9707   0.4734   3.8200   3.3817   5.2861   5.1474   0.4276
 5.7718   4.3280   2.0507   2.2834  14.4425   4.1508   9.9144
 2.1074   4.4184
```

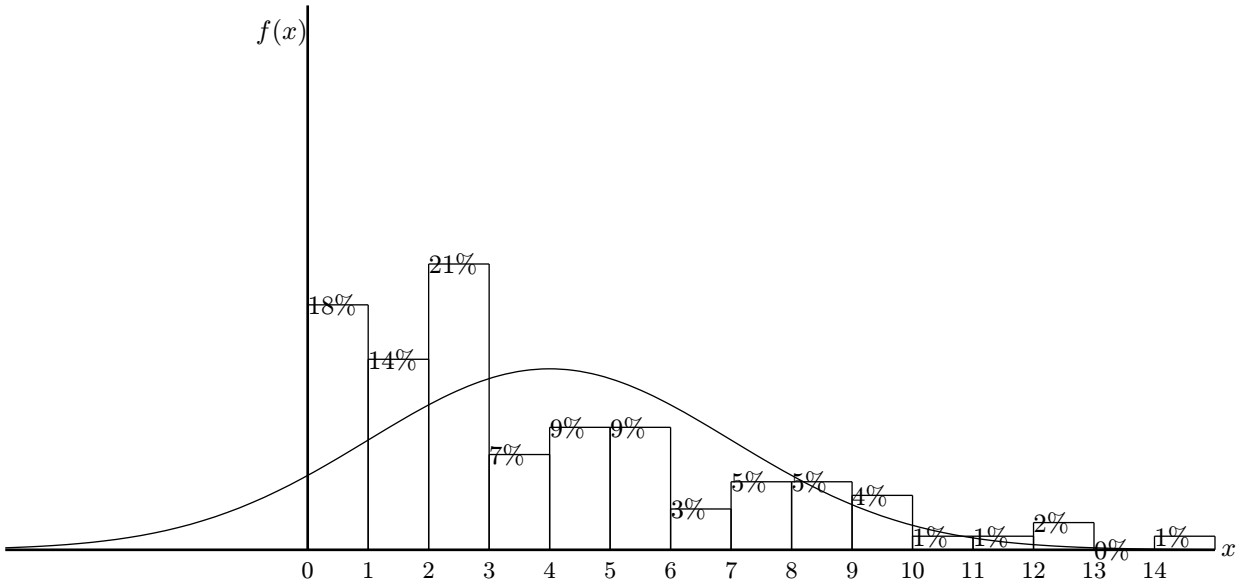
What is the best estimate for the parameter λ? We could eyeball it, as in the following graph. Here, we display a histogram of our data and compare it to a few different exponential curves (of the many that we could draw).



It's a close call by eyeball, but it would appear that a model with $\lambda = .35$ is certainly not the best of these, as it seems to overestimate the actual number of small values of X, and underestimate the number of larger values. The other two values of λ plotted both seem pretty good in different ways. $\lambda = .25$ seems to do better on the left side of the plot, while $\lambda = .15$ seems to do a better job on the right.

Of course, we don't just eyeball it when truly solving this problem. But this is the essence of the problem of parameter estimation - attempting to determine the value of unknown parameters that seems to fit best with the data we collected. In reality, we will rely on techniques like the method of moments or maximum likelihood to solve the problem.

Returning to the first general problem posed in this section - that of model validation - let us say that we had it in our mind (for some reason) that a normal probability model would do a good job with this data. Say we thought $X_1, \ldots, X_n \sim \text{N}(4, 3)$ would describe the data well. We could then use the collected data to see how well our hypothesized model describes it. Below is a similar plot as above, but with the N(4,3) curve drawn over the histogram instead.



We would quickly see that our model was no good; it does a bad job of describing the real world. Of course, in many real situations, it isn't so easy to eyeball it, and we will certainly need more rigorous statistical techniques to answer the question.

2.9.2 Method of Moments

The *method of moments*, or *MOM*, is one method for deriving estimators for unknown parameters. It is a very logical technique, and typically leads to fairly simple computations in real problems.

First, we will look at the simple example of the exponential distribution. This is "simple" because there is only one unknown parameter in the distribution, namely λ. We want to find an estimator that will do a good job of estimating the true value of λ.

The method of moments says to equate the mean for the random variable with the sample mean, then solve for the parameter. For the exponential distribution, we get

$$\frac{1}{\lambda} = \bar{X} \Rightarrow \tilde{\lambda} = \frac{1}{\bar{X}}$$

as our estimator for λ.

Example 2.11: In our protein example, the average of our 100 observations is $\bar{X} = 3.938$. So, the MOM estimator leads to 1/3.938=.2539 as our numerical estimate of λ. Notice this is very close to the middle of the three lines drawn on that plot earlier.

■

For other distributions, such as the normal, gamma, or beta, there are often two parameters (or sometimes more) that need to be estimated. We need to extend the method of moments a bit to handle these. The more generic method involves the following idea. If there are k parameters to be estimated, we will use the first k sample moments, defined as

$$\begin{aligned}
\text{1st sample moment} = \text{sample mean} &= \frac{1}{n}\sum X_i = \bar{X} \equiv m_1 \\
\text{2nd sample moment} &= \frac{1}{n}\sum X_i^2 \equiv m_2 \\
\text{3rd sample moment} &= \frac{1}{n}\sum X_i^3 \equiv m_3 \\
&\vdots \\
k\text{th sample moment} &= \frac{1}{n}\sum X_i^k \equiv m_k
\end{aligned}$$

and the first k population moments which are defined as

$$\begin{aligned}
\text{1st population moment} &= \mu_X \\
\text{2nd population moment} &= \mu_{X^2} \\
\text{3rd population moment} &= \mu_{X^3} \\
&\vdots \\
k\text{th population moment} &= \mu_{X^k}
\end{aligned}$$

These population moments after the second are not necessarily always easy to find, but fortunately we don't often have more than two parameters, so it is only the first two we need to deal with.

The next step is to equate sample moments with population moments of the same order. This creates k equations with k unknown parameters. Solving these equations gives us the simultaneous method of moments estimators for those parameters. This is best shown through an example.

Let's look at the gamma distribution. Let $X_1, \ldots, X_n \sim \text{Gamma}(\alpha, \beta)$. Both α and β are unknown parameters and we want to use the method of moments to find estimators for them. In fact, as we'll see, other methods such as maximum likelihood method, when applied to the gamma distribution, will not find closed-form solutions for estimators of α and β. So, the MOM is very useful here.

First, we'll refer to the first two sample moments as m_1 and m_2 to start with so there isn't so much notation floating around. Now, the first population moment is just $\mu_X = \alpha\beta$ as mentioned in Section 2.7.2. The second population moment takes just a little more work; we need to twist around the variance formula to calculate μ_{X^2} and plug in the known variance of the gamma distribution:

$$\sigma_X^2 = \mu_{X^2} - \mu_X^2$$
$$\alpha\beta^2 = \mu_{X^2} - \alpha^2\beta^2$$

leaving us with

$$\mu_{X^2} = \alpha\beta^2 + \alpha^2\beta^2 = \alpha(\alpha+1)\beta^2$$

.

So our system of two equations is now

$$\alpha\beta = m_1$$
$$\alpha(\alpha+1)\beta^2 = m_2$$

and we need to solve this for α and β. We can simply rearrange the first equation to get $\beta = m_1/\alpha$ and make this substitution in the second equation and solve for α. After some algebra, we get

$$\alpha = \frac{m_1^2}{m_2 - m_1^2}$$

Plugging this back into the equation for β gives

$$\beta = \frac{m_2 - m_1^2}{m_1}$$

Now we can substitute in for m_1 and m_2 to see our final MOM estimators for the parameters are

$$\tilde{\alpha} = \frac{\bar{X}^2}{\frac{1}{n}\sum X_i^2 - \bar{X}^2}$$

$$\tilde{\beta} = \frac{\frac{1}{n}\sum X_i^2 - \bar{X}^2}{\bar{X}}$$

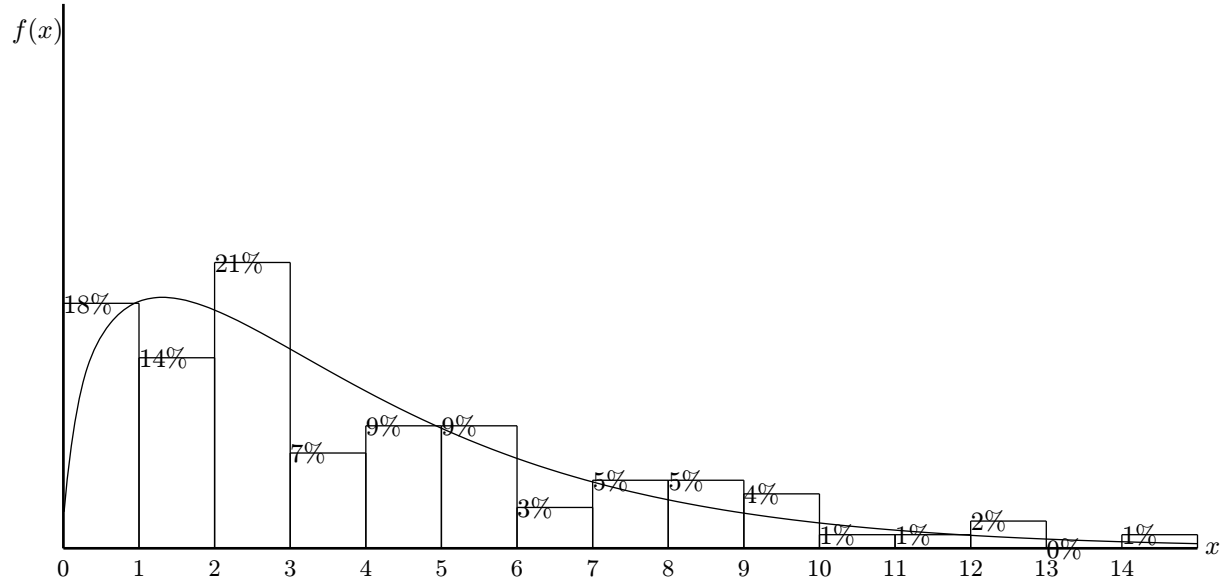
Example 2.12: Let's go back to the protein degradation example. Suppose we started out modeling our data using a generic gamma distribution instead of exponential (which is just a special case of the gamma where we set $\alpha = 1$).

So we have $X_1, \ldots, X_n \sim \text{Gamma}(\alpha, \beta)$ and will use the equations above as estimators for these parameters. All we need to do is calculate $\bar{X}$ and $\frac{1}{n}\sum X_i^2$ from the 100 observations. We have seen already $\bar{X} = 3.938$ and a computer package or calculator will show us that $\frac{1}{n}\sum X_i^2 = \frac{1}{100}(2586.945) = 25.869$. So our estimates using the MOM estimators are

$$\tilde{\alpha} = \frac{(3.938)^2}{25.869 - (3.938)^2} = 1.4967$$

$$\tilde{\beta} = \frac{25.869 - (3.938)^2}{3.938} = 2.6311$$

We notice that our estimate of α is 1.4967, a bit more than the $\alpha = 1$ we assumed by using the exponential distribution previously. Let's see how these estimates do visually with our data:



Not too bad a fit. The gamma distribution helps recognize the slight bump in the number of observed values between 2 and 3. We would probably conclude that this is a better probability model for this data, at least by sight.

■

You should double check the result above for the gamma distribution, and try the MOM yourself for other distributions, such as the normal.

2.9.3 Maximum Likelihood Estimation

Maximum likelihood estimation, or *MLE*, is a more rigorous procedure for finding estimators for parameters than is the method of moments. Although many times we will get the same answer whether we use the MLE or MOM methods, sometimes it can be different, particularly for more complex probability models. In such cases, the MLE-based estimator is often considered the better one. The tradeoff is that finding the maximum likelihood estimator can be more difficult, and sometimes even require computational effort.

The method of maximum likelihood proceeds as follows. (Actually, in real situations this procedure will not work correctly, but we won't discuss those here.)

1. Write down the joint pdf of the random sample. We will also refer to this as the *likelihood function* of the parameter(s), and view it as a function of the parameter(s) instead of a function of the random variables. Call this function $L(\theta)$ where θ is our generic symbol for the parameter of the model.
2. Take the natural log of L, giving us $\ln L(\theta)$.
3. Differentiate with respect to each parameter individually (or just once if there is only one parameter).
4. Finally, set the resulting equation(s) equal to zero and solve for the parameters. The result gives you the MLE for these parameters (other than one other technical step that we won't discuss). This last step can be impossible in some situations.

The same procedure applied whether we have discrete or continuous variables.

Let's take a look at an example using the exponential distribution, which is easy to deal with. We have $X_1, \ldots, X_n \sim \text{Exponential}(\lambda)$ and so the likelihood function for λ, and therefore the joint pdf of $X_1, \ldots, X_n$ is (which we, as before, can easily write down because of the independence rule)

$$\begin{aligned} \mathrm{L}(\lambda) = f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) &= f_{X_1}(x_1) \times \cdots \times f_{X_n}(x_n) \\ &= \left(\lambda e^{-\lambda x_1}\right) \cdots \left(\lambda e^{-\lambda x_n}\right) \\ &= \lambda^n e^{-\lambda \sum x_i} \end{aligned}$$

Continuing, we have

$$\ln \mathrm{L}(\lambda) = n \ln \lambda - \lambda \sum x_i$$

and differentiating and setting equal to 0 we get

$$\frac{d \ln \mathrm{L}(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum x_i = 0$$

Now, solving this for λ gives us $\hat{\lambda} = 1/\bar{X}$ as the maximum likelihood estimator of λ. Notice this is the same as the MOM estimator, and also is one that makes sense since λ is 1 over the mean of the distribution.

Try using the maximum likelihood method for the gamma distribution on your own. Since there are two unknown parameters, you will take two derivatives, and have two equations set equal to zero. You will notice that for the gamma distribution, these equations cannot be solved analytically for the parameters. We would have to rely on computational techniques to solve them numerically.

2.9.4 Confidence Intervals

A *confidence interval*, or *CI*, is a way of estimating a parameter by giving a range of likely values for the parameter. In examples in the previous section, we provided an estimate for an unknown parameter simply

by reporting a single number. This is convenient and easy to interpret, but lacks some important information. In fact, we know that the single number provided as an estimate is not correct. So, in addition to that single number, we would like some information regarding how precise that number is as an estimate of the parameter. This is what a confidence interval does.

Say we have an unknown parameter θ in a problem. A typical way we might report a confidence interval for θ (after having collected and analyzed some data) is to say "a 95% confidence interval for θ is (4.56, 5.78)." The 95% is our *confidence level*. It shows how sure we are of the range we are reporting. The interval itself, 4.56 up to 5.78 in this example, is the range of values in which the data suggests θ really falls.

Computing a confidence interval is quite easy in some cases, and can be much more difficult in others. For this course, we will take a simple view of confidence interval calculations (unless otherwise noted). We will use the following formula to compute a CI for a parameter θ:

$$95\%\ \text{CI} = \hat{\theta} \pm 2\ \sigma_{\hat{\theta}} \tag{2.9}$$

where $\hat{\theta}$ is the estimate of θ based on some method like MOM or MLE, and $\sigma_{\hat{\theta}}$ is the standard deviation of that estimator.

The use of this formula is fine for many situations. Essentially, it is based on the fact that maximum likelihood estimators approximately have a normal distribution when the sample size is relatively large. So, if we base estimators on the MLE method and have a somewhat large sample size, this CI formula will work just fine for our purposes.

Note, if you want a 99% CI instead, replace the value "2" with "2.576".

2.10 Hypothesis Testing

Hypothesis testing, like parameter estimation, is an inferential procedure in statistics. The situation is that we have a probability model with unknown parameters. Instead of simply estimating values for the parameter(s), we might want to test hypotheses about it/them. This is the topic of hypothesis testing. It turns out that parameter estimation and hypothesis testing are very closely related.

2.10.1 General Setup of a Hypothesis Test

A hypothesis test always consists of two hypotheses, called the null hypothesis and the alternative hypothesis. The null is labeled H_0 and the alternative is labeled H_a . Typically, the null is the hypothesis that we will believe by default, unless our data provides strong evidence that we should instead believe the alternative. From a probability model standpoint, however, the null hypothesis is typically the one with a simpler probability

model, and the alternative is the one with a more complex model (such as having more unknown parameters).

To help understand the terminology as we discuss it, let's refer to a common hypothesis testing example in statistical genetics. Consider a locus with two alleles A and a, whose population proportions are P_A and $P_a = 1 - P_A$. The three genotypes at this locus are AA, Aa, and aa, with probabilities P_{AA}, P_{Aa}, and P_{aa}, which sum to 1.

A common question is whether this locus is in Hardy-Weinberg Equilibrium (HWE), a situation we will discuss in much more length in Chapter 3. HWE says that the relationship between genotypic and allele frequencies has the following form

$$\begin{aligned} P_{AA} &= P_A^2 \\ P_{Aa} &= 2P_A(1 - P_A) \\ P_{aa} &= (1 - P_A)^2. \end{aligned}$$

The probability model that we would be employing if HWE was not the case would be

$$(X_{AA}, X_{Aa}, X_{aa}) \sim \text{Multinomial}(n, P_{AA}, P_{Aa}, P_{aa}) \tag{2.10}$$

and the model would reduce down to the following if HWE was true:

$$(X_{AA}, X_{Aa}, X_{aa}) \sim \text{Multinomial}(n, P_A^2, 2P_A(1 - P_A), (1 - P_A)^2) \tag{2.11}$$

So, our hypothesis comes down to a question of which probability model best describes the data. A way to write the two hypotheses would be

$$\text{H}_0 : \begin{cases} P_{AA} &= P_A^2 \\ P_{Aa} &= 2P_A(1 - P_A) \\ P_{aa} &= (1 - P_A)^2 \end{cases}$$

H_a :The above relationships do not hold

Or we might say that our null hypothesis is that Equation (2.11) is the correct probability model and the alternative hypothesis is that Equation (2.10) is the correct probability model.

Now, we ultimately want to make some statement about which hypothesis we believe. This, of course, will be based on data that we collect. Generally speaking, we will let the data decide for us: which hypothesis does it most support? The only caveat to this is that, as stated above, we will only claim that it supports the alternative hypothesis if there is very strong evidence in that direction; otherwise, by default, we will believe the null.

2.10.2 The Decision Process

Making that decision is often done through a hypothesis testing process referred to as the Neyman-Pearson (NP) method of hypothesis testing. Although there are a number of different general methodologies for conducting tests that exist, the NP method is by far the most common and well-known. We will focus on that methodology.

The NP method of performing a hypothesis test consists of the following steps. We will assume that the data we have collected is represented by the random variables $X_1, \ldots, X_n$.

First, we must determine an appropriate *test statistic* to use. A test statistic is a function of the data, and it summarizes the data in a way that will be useful for making our decision. It is very similar in nature to an estimator when we discussed parameter estimation. An estimator was a statistic (a function of the data) that summarized the data in such a way that it provided a reasonable estimate for a parameter.

The choice of an appropriate test statistic is a problem that involves some statistical theory that we will not be going into (as opposed to finding parameter estimators - where we did discuss the theory of MLE's, for example, to choose appropriate estimators). For particular situations, we will just mention what the appropriate statistic is.

The important thing we need to know about the test statistic we choose is what probability distribution it has *when the null hypothesis is true*. Remember that a test statistic, just like an estimator, is itself a random variable. Its calculated value depends on the exact sample selection and other randomness involved in the data collection process. So the value we would get for one sample of size n would not be the same as the value we would get for a different sample of size n.

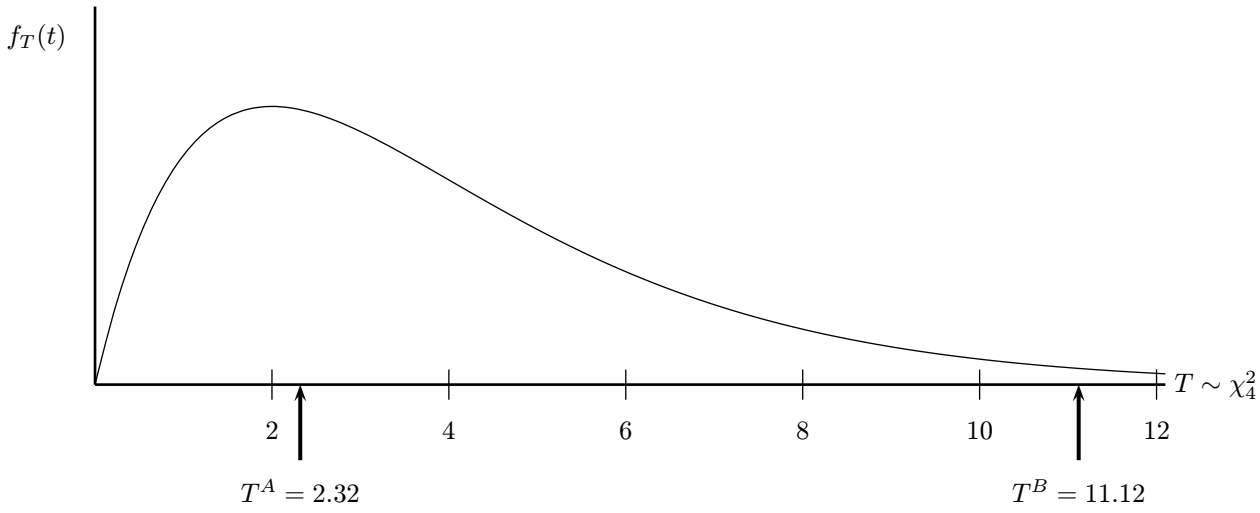
Therefore, the probability distribution of the test statistic describes the likelihood of different values being calculated after collecting data. Keep in mind, that this distribution is the distribution we would get *when the null hypothesis is true*. This is a key fact that we'll use to help make a decision about our hypotheses.

Now that we have a test statistic and know its distribution under the null, we can calculate the value of the test statistic based on our data. Our decision process comes down to the following general concepts:

- If the calculated value of the test statistic seems very unusual based on its distribution (remember this is the distribution as if the null were true), that will be a sign to us that the null hypothesis is actually not true. We will say that we "reject the null hypothesis."
- On the other hand, if the calculated value of the test statistic is right in line with typical and common values of its distribution under the null, that will be a sign to us that the null hypothesis is a reasonable model, or at least that there isn't evidence against it. We will say that we "do not reject the null hypothesis."

Let's set up a hypothetical example to help see this clearly. Say we have decided on some test statistic -

call it T. T might be $\bar{X}$ or some other function of the data we have collected. Let's also say that we know that $T \sim \chi^2_4$ if the null hypothesis were true. (Note: We haven't called this statistic T because it has a t-distribution; the T stands for Test Statistic). So, we can draw a picture of this distribution



Also in this picture are the values calculated for T for hypothetical datasets we might have (dataset A and dataset B). Notice the value of T calculated for dataset A (2.32) would be considered a very "typical" or "common" value to observe for T if the null were true, because it falls right in the "high probability" area of its distribution. So, we are not surprised that this value was observed, and in fact it seems to support the notion that the null hypothesis really is true.

However, now notice that value of T calculated for dataset B (11.12). This would certainly not be considered typical or common to observe as a value for this distribution. It is well out into the right tail of the distribution - a very low probability area of the curve. So, since this is the distribution of test statistic values we would expect if the null were true, but we actually calculated a value very atypical of this distribution, logic would suggest that our assumption that the null was true is probably wrong. Therefore, our data tends to support the notion that the null is not true, and the alternative is.

Practically speaking, we would like to have a little more mathematical rigor behind our assessment of whether the test statistic seems very unusual or somewhat typical to have calculated if the null were true. This is provided by the calculation of the *p-value* of the test. The definition of the p-value is

$$\begin{aligned}\text{p-value} = \text{P(The test statistic could be even more extreme than} \\ \text{its calculated value, according to its distribution under the null)}\end{aligned} \tag{2.12}$$

First of all, we notice the p-value is a probability. Its value will be between 0 and 1. Second, it measures how unusual, or atypical, it was to have calculated that value of the test statistic if the null hypothesis actually were true. Smaller values (close to 0) indicate a test statistic that was somewhat unusual. Larger values (further from 0) indicate a test statistic that was more in line with the null and fell right in the "meat" of

the distribution.

Typically, we might use a cutoff of, say, .05 or .01, to make a final determination of whether the test statistic was unusual or not. This cutoff is called the *significance level of the test* and is referred to as α in notation. Our rule, then, might be:

- If the p-value is less than or equal to α, we will reject the null hypothesis (the data provided strong evidence against it), or
- if the p-value is greater than α, we will not reject the null hypothesis (the data did not provide strong enough evidence against it).

So, we can use the p-value to make a clear reject or non-reject decision, or we can simply use it as a measure of how strong our evidence was against the null, and let others decide for themselves how they want to interpret it.

Calculating the p-value also needs to be discussed. The calculation is a probability calculation, and so it will be based on the distribution of the test statistic. Typically, these calculations will be done on the computer using Excel functions mentioned earlier, or other software or websites.

For example, let's return to our hypothetical situation above with a test statistic $T \sim \chi^2_4$. Consider dataset B, where we calculated $T = 11.12$. The p-value of this test is, by the definition in Equation (2.12)

$$\text{p-value} = \text{P}(T \geq 11.12)$$

We can compute this using Excel's cdf function for the chi-square distribution.

Excel Formula

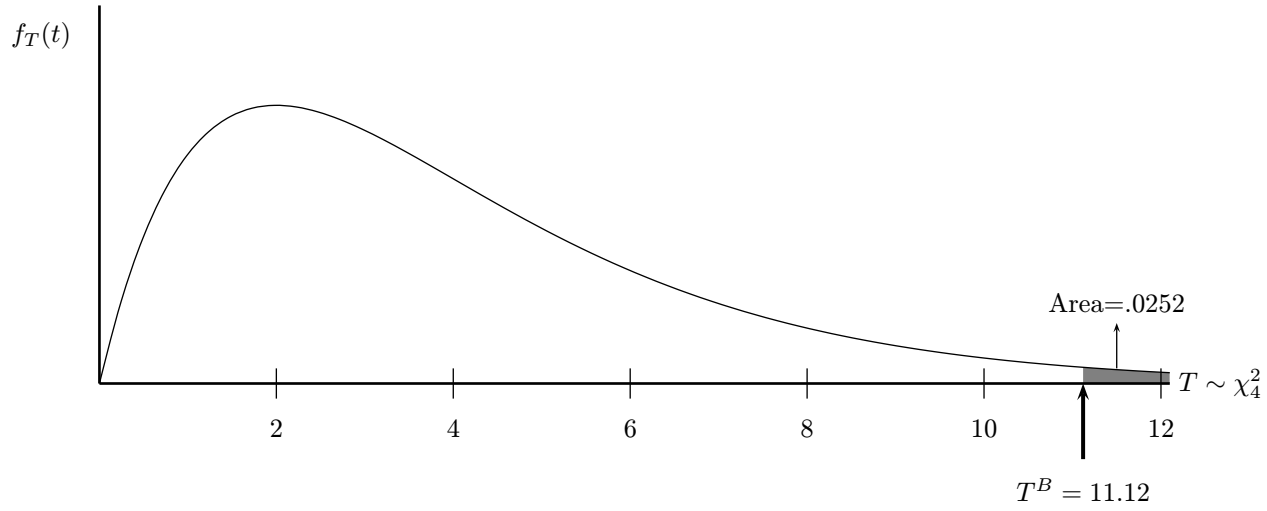
```
CHIDIST(11.12,4) = .0252
```

So,

$$\text{p-value} = \text{P}(T \geq 11.12) = .0252.$$

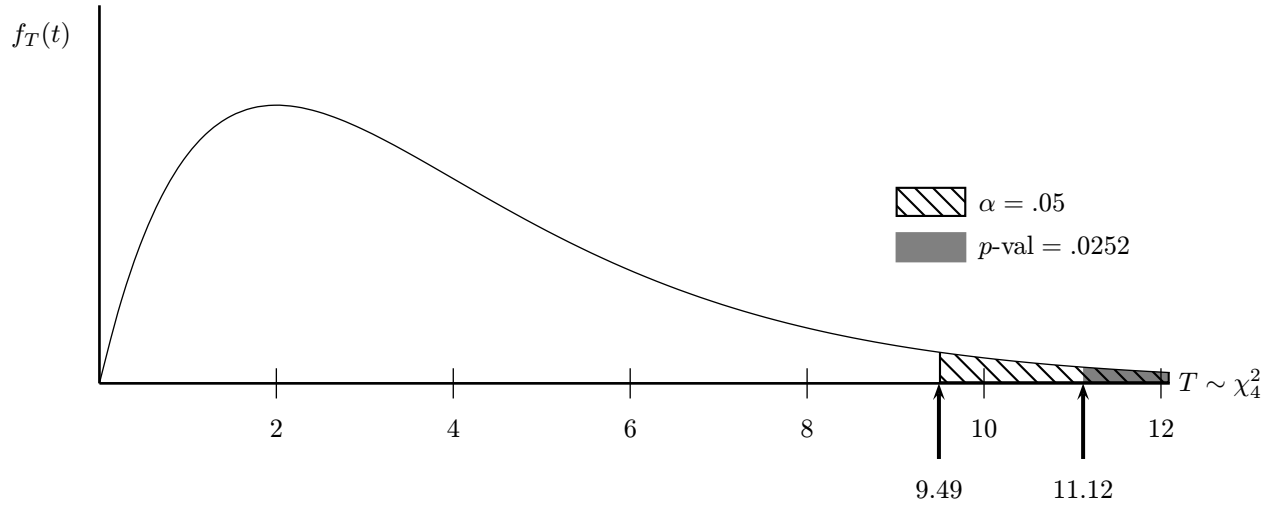
According to our above rule, if our α was set to be .05, we would reject the null hypothesis in the favor of the alternative. In other words, the evidence against the null was strong enough to not believe it is true.

Graphically, this p-value is the area under the pdf of T to the right of 11.12.



This also helps to show that the more evidence there is against the null hypothesis, the smaller and smaller the p-value will be (closer to 0). This is because [More evidence against H_0] $\rightarrow$ [Test statistic more atypical for its distribution] $\rightarrow$ [Area to the right of test statistic small] $\rightarrow$ [p-value small].

Notice the role α plays in this. Let's take the previous picture, but also plot the area $\alpha = .05$ in the right tail of the distribution.



Notice that essentially, the value chosen for α determines where the line is drawn that separates "typical values" of the distribution of T (values that would lead to not rejecting H_0) from "unusual values" of the distribution of T (values that lead to rejecting H_0). These unusual values, those to the right of the line, are called the *rejection region.*

The value of the test statistic that falls where the cutoff is made is called the *critical value* of the test. In the above picture, that value is 9.49. We'd say that we would reject H_0 for values of T greater than 9.49. Notice that $9.49 = F_T^{-1}(.95)$; it is the 95th percentile of the distribution of T. It can be found in Excel by

Excel Formula

```
CHIINV(.05,4) = 9.4877
```

Finally, let's also calculate the p-value associated with dataset A in the above example. We would use Excel to calculate

Excel Formula

```
CHIDIST(2.32,4) = .6771
```

Therefore, .6771 is the p-value associated with that dataset and that value of the test statistic. It would be considered quite a large p-value, and therefore a sign that the data provides no evidence against the null. In fact, as we've discussed, that data is perfectly in line with what we would expect if the null is true.

In the above example, we defined "more extreme than its calculated value" as being values further to the right. In some testing situations (and in our above example), that is the correct way to look at it. But in others, the p-value might be calculated by considering "more extreme" to mean "further to the left", or in some cases to mean "further out in either tail" of the distribution. Which meaning to use will be discussed at the appropriate time, and often will be clear from the setup of the particular test we are conducting.

2.10.3 Errors in Hypothesis Testing

It is important to recognize that whenever we conduct a hypothesis test, there is a chance that our conclusion will be wrong. Table 2.4 shows the situations we may be in when we conduct a hypothesis test. The columns represent the reality of the null hypothesis: it is either really true or really false. Remember, the null and alternative hypotheses make statements about populations (about the underlying truth of real-world phenomena), and we will never really know which is really the truthful statement. We can only base decisions on a representation of that population, that is, on the data we collect. Despite that, one and only one, of these hypotheses is *actually* true.

The two rows represent the two possible decisions we can make: to either believe the null hypothesis (i.e., not reject it), or to not believe it (i.e., reject it). Depending on which row and column our situation falls in, we have either made a correct decision or an incorrect decision. This table summarizes the possibilities:

We can see the two situations that errors will be made: if H_0 really is true, but our data incorrectly leads us to reject it (called a Type I Error), and if H_0 really is false, but our data incorrectly leads us to not reject it (called a Type II Error). We, of course, would like to not make either type of error, but the possibility is a fact we have to live with since we only have partial information (i.e., a sample of data). So, the best we can do is to make the probabilities of these errors small. We do so in the following way.

Table 2.4: Errors in hypothesis tests

	Truth of H_0	
Decision	H_0 True	H_0 False
Not Reject H_0	Correct	Type II Error
Reject H_0	Type I Error	Correct

First let's discuss Type I Error. The probability of making a Type I Error is α, or the significance level of the test - a quantity that we have already come across. In other words, the α value we chose in the last section actually represents an error probability. Clearly we want this to be low, so we will typically want to choose α to be values such as .01, .05, or .10. We would decide to make it smaller (say, .01 or even lower) if it is very important that we do not make a Type I Error. In many testing situations, a Type I Error represents an error we want to be careful to avoid. For example, in a medical scenario, a Type I Error might be believing that a certain surgical procedure is useful (rejecting H_0) when actually it isn't (H_0 was actually true). We'd like a small chance of having this be the outcome, so would probably set α quite low.

Notice that for a Type I Error, the Neyman-Pearson approach to hypothesis testing allows us to set its probability ahead of time, giving us the opportunity to make it as low as we need it to be. It is controllable.

However, the tradeoff is that we cannot directly control the probability of a Type II Error. This probability is often referred to as β. We cannot specify as part of the testing procedure how low we want β to be. This is unfortunate, but we can *indirectly* control β through a couple of means:

- β will be smaller if we use an appropriate testing procedure. This refers to the fact that there are often different ways to handle any particular test we want to conduct. In other words, there are different test statistics we could use (just as in parameter estimation, there are different estimators we could use). Choosing the "best" one will serve to make β smaller than if we don't.
- β can generally be made smaller by taking a larger sample. The tradeoff here is, of course, that a larger sample requires more work, so we won't have unlimited resources to continue decreasing β this way.

Regarding the second point above, it is often possible to determine prior to collecting data how large a sample we need in order to make β have a certain value. In practice, these calculations are almost always made because it allows the researcher to plan the size of the study and assure him or herself of having a the probability of Type II Error be as low as they want.

It should be noted that the tradeoff between sample size and β can be such that it may not be possible to make β be a small number; the researcher may have to live with β being as large as .10, .20, or often even much higher. This fact is the reason that the null and alternative hypotheses are setup as we discussed. The null is the one we will believe by default, and the alternative is the one that we won't believe unless there is strong evidence in favor of it. In this setup, a Type I Error (whose probability we can control) is the more

dangerous error to make, and a Type II error (whose probability is not directly controllable) is unwanted but not as dangerous.

2.11 Chi-squared Goodness-of-Fit Test

The first specific type of test we will discuss is one that can be used to formally test whether the probability model chosen is appropriate or not. It essentially does, in a more mathematically formal fashion, what we did by eyeball when comparing histograms to pdf curves in Section 2.9.1. The test procedure that is often used for this type of problem is called a *chi-squared goodness-of-fit test*.

2.11.1 Model is Completely Specified

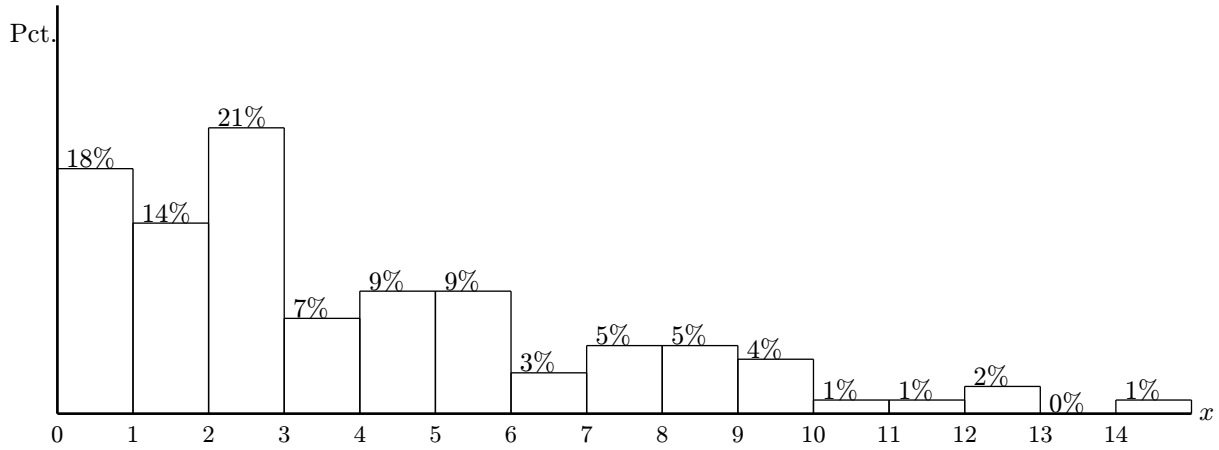
The first situation we will consider is when we have completely specified the probability model we are suggesting in the null hypothesis. By "completely specified", we mean that there are no unknown parameters in the model. There is one, and only one, probability distribution that has been specified for the data.

As an example, let us return to one of the protein degradation models we looked at in Section 2.9.1. We wrote our probability model as $X_1, \ldots, X_n \sim \text{N}(4, 3)$. Now, we want to formally test whether this is the correct model. Our hypotheses are

$$\text{H}_0 : \text{N}(4, 3) \text{ is an appropriate model for this data}$$
$$\text{H}_a : \text{N}(4, 3) \text{ is not an appropriate model for this data}$$

Notice the null hypothesis represents only one possible distribution - the Normal distribution with mean 4 and standard deviation 3. There are no unknown parameters in this distribution.

Let's review the data from that problem. Here was the histogram:



Since the sample size was $n = 100$, the percentages in each bar (e.g., 18% in the 1st bar) are also the number of observations that fell in that bar (e.g., 18 observations were between 0 and 1).

How might we conduct this test? The general approach for any hypothesis test, as we have discussed, is that we first need to decide on a test statistic to use (i.e., an appropriate way to summarize the dataset) and know what distribution that statistic has. Let's consider the following approach to summarizing the data in preparation for conducting this test.

First, group the continuous data into k logical discrete categories. There are some rules of thumb we need to follow when doing this:

(a) Each category should represent a contiguous range of values. In other words, one category should not represent the ranges [0,1) AND [2,3) since these are not contiguous.

(b) Every possible value of the random variable must fall into one, and only one, category. In other words, categories should not overlap, and together must comprise the entire range of the random variable. Note: this last part should not be interpreted as "the entire range of the data" or as "all the observed values in the dataset." The random variable in the probability model is likely to have a range greater than the observed data, as noticed in our example (a normal random variable can have any real number value, but the data ranged only from 0.0426 to 14.4425).

(c) Categories should be created so that the number of observations in any category is "not too small." We will discuss the reason for this, and a more formal definition of "too small" later.

Using these rules, let's consider the following categorization for our example.

Category	Range of Values	Observed Probability	Observed Count
1	$(-\infty, 1)$	18%	18
2	[1,2)	14%	14
3	[2,3)	21%	21
4	[3,4)	7%	7
5	[4,5)	9%	9
6	[5,6)	9%	9
7	[6,7)	3%	3
8	[7,8)	5%	5
9	[8,9)	5%	5
10	$[9, \infty)$	9%	9

This meets all of our rules, although for rule (c), we'll just assume it does for now. Notice the third and fourth columns in this table are labeled "Observed Probability" and "Observed Count". These are the percentage of observations and the total number of observations, respectively, that fell into each category in the data we collected. The total in these columns should be exactly one and exactly the sample size, respectively.

Now, we need to calculate *expected* percentages and counts for each category. The expected percentage for a category is the percentage of observations that we would expect to fall in the category if the null hypothesis were true. For example, the expected percentage for category 7 in the table is the probability that a N(4,3) random variable will fall between the values 6 and 7. This is $\mathrm{P}(6 \leq X < 7) = F_X(7) - F_X(6)$. The Excel calculation is

Excel Formula

```
NORMDIST(7,4,3,TRUE)-NORMDIST(6,4,3,TRUE)=.0938
```

The expected count for a category is the sample size multiplied by the expected probability. In the case of category 7, this is $100 \times .0938 = 9.38$. After making these calculations for each category, we can put the results in two new columns in our table:

Category	Range	Observed Prob	Observed Count	Expected Prob	Expected Count
1	$(-\infty, 1)$	.18	18	.1587	15.87
2	[1,2)	.14	14	.0938	9.38
3	[2,3)	.21	21	.1169	11.69
4	[3,4)	.07	7	.1306	13.06
5	[4,5)	.09	9	.1306	13.06
6	[5,6)	.09	9	.1169	11.69
7	[6,7)	.03	3	.0938	9.38
8	[7,8)	.05	5	.0674	6.74
9	[8,9)	.05	5	.0434	4.34
10	$[9, \infty)$	.09	9	.0478	4.78
Total		1.00	100	1.00	100

The concept behind the goodness-of-fit test now says: "Compare the counts observed for each category to the counts we would expect in each category if the null hypothesis were true. If these counts are relatively close to each other, category by category, then that is evidence that the null is true. If they are generally far apart, then that is evidence that the null is false."

This gives some intuition behind the test statistic that we use, which we'll now define. It is

$$T = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j} \tag{2.13}$$

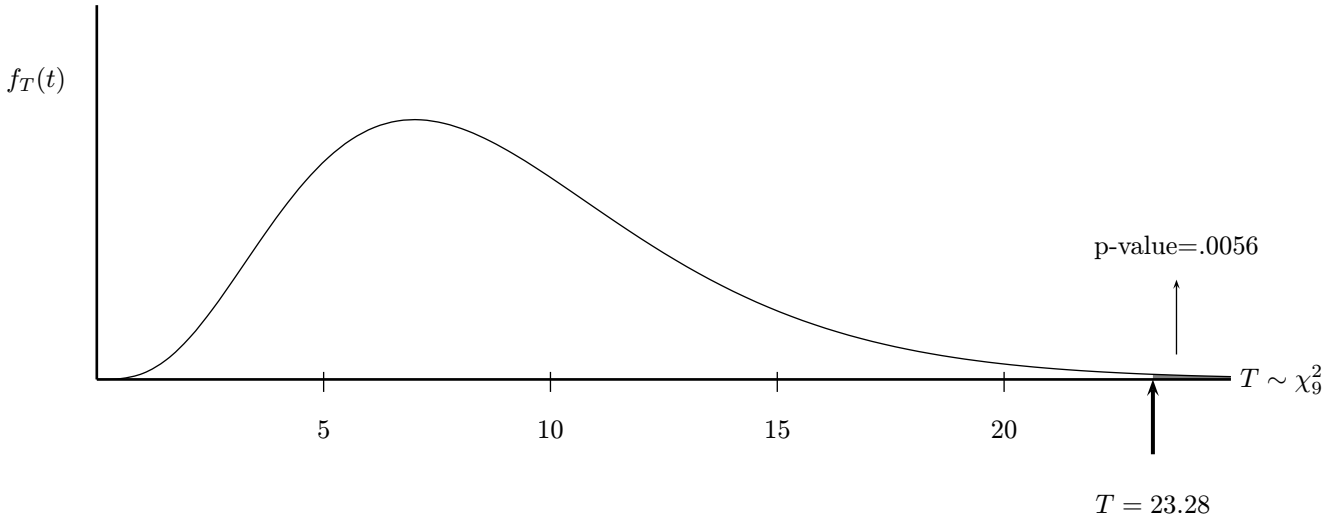
where the sum is over all k categories defined, O_j is the observed count for category j, and E_j is the expected count for category j. Notice that the numerator will be larger the further apart the observed and expected counts are, and will be smaller the closer together they are. Our logic then suggests that larger values of T will lead to smaller p-values - a sign of more evidence against H_0 .

It turns out that $T \sim \chi^2_{k-1}$ if the null hypothesis is true. That it, it has a chi-squared distribution with degrees of freedom equal to one less than the number of categories used. We'll use this distribution to calculate p-values as a gauge as to how strong our evidence is against H_0 .

Let's follow the above example through. We've set up all the necessary information in the above table. We now just need to calculate the value of the test statistic and use that to calculate the resulting p-value. The test statistic calculation is (by applying Formula (2.13) to each row of the table)

$$T = \frac{(18 - 15.87)^2}{15.87} + \frac{(14 - 9.38)^2}{9.38} + \cdots + \frac{(9 - 4.78)^2}{4.78} = 23.27$$

For our problem, $T \sim \chi^2_9$ if H_0 true. A picture of the result is



As seen in the picture, the p-value calculation is p-value=P($T > 23.28$) $= 1 - F_T(23.28) = .0056$. The calculation in Excel is

Excel Formula

```
CHIDIST(23.28,9)=0.0056
```

The p-value is quite small, certainly smaller than any significance level typically chosen. It appears that with strong evidence, we can reject the null hypothesis and conclude that a N(4,3) probability model does not fit this data well. This is not a surprising conclusion considering the picture we analyzed in the previous chapter.

A final note on the terminology used for this test procedure: the "goodness-of-fit" test. It should be a little more clear now why that terminology is used. We are trying to determine how well (or "good") a specified distribution "fits" with the data we observed.

2.11.2 Model is not Completely Specified

A slightly different take on the test discussed above is as follows. In some (in fact, many) situations, we might know or be willing to specify the general form of the distribution (such as normal, gamma, exponential, etc.), but not know the values of the parameters for whichever distribution we specify. In the above example, not only did we hypothesize that the data followed a normal distribution, but we were very specific about which normal distribution: the normal distribution with mean 4 and standard deviation 3. Based on our discussions in previous chapters, it should be no surprise that in most situations we don't know, or are unwilling to specify, the exact parameters of our model. The null hypothesis of the preceding section was somewhat limiting in this sense.

Let's take that example and put a different spin on it. Now let's say that we hypothesize that the data follows a gamma distribution, but are unable or unwilling to specify the exact parameters of the distribution. In other words, we believe that some form of the gamma distribution may fit the data well, but are not sure exactly which form.

Notice that in many respects, this is similar to the estimation problem we had discussed. We modeled the data with a gamma distribution, but did not specify the parameters. We then estimated the parameters based on the data we observed. The slight difference is that before, we estimated the parameters assuming the data had a gamma distribution, but made no statement about whether the gamma distribution itself made sense. Now, we will take that extra step and decide whether the entire model is sensible.

Now, the hypotheses we will test are stated as follows:

H_0 :Gamma(α, β) is an appropriate model for this data

H_a :Gamma(α, β) is not an appropriate model for this data

The procedure we use is still referred to as the chi-square goodness-of-fit test. And we follow the same general procedure as discussed in the previous section, with just two amendments:

(1) Since we don't know the values of the parameters, we must first estimate them. Typically maximum likelihood estimates would be used, but method of moments-based estimates are ok substitutes if necessary. Then, we base expected probability and expected count calculations on these estimated values.

(2) We still use the same test statistic formula, and it still has a chi-squared distribution. However, the degrees of freedom is now ($k - 1-$ Number of Estimated Parameters).

Let's make the calculations for our example. First, we have already calculated estimates, based on the method of moments, for these parameters in Section 2.9.2. They were $\hat{\alpha} = 1.4967$ and $\hat{\beta} = 2.6311$. Using these and Excel to compute probabilities for a gamma distribution, we get the following table of observed and expected counts (and probabilities).

Category	Range	Observed Prob	Observed Count	Expected Prob	Expected Count
1	$(-\infty, 1)$	.18	18	.1418	14.18
2	[1,2)	.14	14	.1817	18.17
3	[2,3)	.21	21	.1614	16.14
4	[3,4)	.07	7	.1308	13.08
5	[4,5)	.09	9	.1014	10.14
6	[5,6)	.09	9	.0767	7.67
7	[6,7)	.03	3	.0570	5.70
8	[7,8)	.05	5	.0419	4.19
9	[8,9)	.05	5	.0305	3.05
10	$[9, \infty)$	.09	9	.0768	7.68
Total		1.00	100	1.00	100

This leads to the following calculations for T and the p-value

$$T = \frac{(18 - 14.18)^2}{14.18} + \frac{(14 - 18.17)^2}{18.17} + \cdots + \frac{(9 - 7.68)^2}{7.68} = 9.55$$

Now, since we had to estimate two parameters before making the expected count calculations, we have $T \sim \chi^2_{10-1-2} = \chi^2_7$ if H_0 true. The p-value is

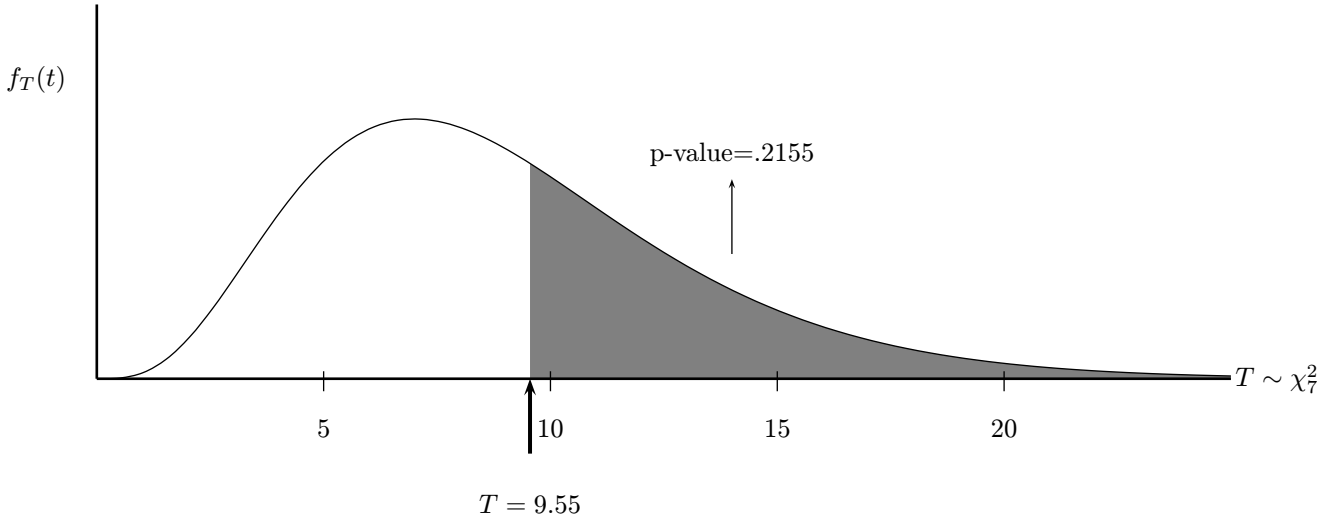
$$\text{p-value} = P(T > 9.55) = 1 - F_T(9.55) = 0.2155.$$

In Excel this is done by

Excel Formula

```
CHIDIST(9.55,7)=0.2155
```

This leads to the following picture of our test statistic



We can see that this calculated value of T from our data is not unusual at all under the null hypothesis, leading to the fairly high p-value. This would lead us to not reject H_0 , and conclude that the gamma model does do a reasonably good job of describing this phenomena. (More formally, we would say that we DO NOT have enough evidence to claim that the gamma model is NOT correct).

2.11.3 Goodness-of-Fit Test for a Discrete Distribution

The discussion of goodness-of-fit tests so far has focused on situations where our probability model is a continuous distribution. But there is no reason we can't easily extend this procedure to discrete distributions. There only difference is that the categories we construct are no longer continuous ranges of real numbers, but groupings of discrete values. In fact, many categories will simply be represented by a single integer, as we'll see.

As an example, let's consider the situation where we observe the number of crossover events that occur between two loci on a chromosome. We make 85 such observations. We will model this data using a Poisson distribution, so we have $X_1, \ldots, X_{85} \sim \text{Poisson}(\lambda)$. The resulting data are

```
0  1  0  2  0  0  2  0  0  1  0  0  0  0  0  0  1  0
```

1	2	0	2	1	1	2	3	0	1	1	2	0	1	1	1	1	0
2	2	1	1	2	0	1	0	0	2	1	0	2	0	0	0	1	0
0	3	1	1	0	0	0	0	0	0	0	0	0	1	0	2	1	1
1	1	0	0	0	0	0	1	1	0	1	2	0					

Is the Poisson model appropriate? Our hypotheses are

$$\begin{aligned} &\mathrm{H}_0 \text{ :Poisson}(\lambda) \text{ is an appropriate model for this data} \\ &\mathrm{H}_a \text{ :Poisson}(\lambda) \text{ is not an appropriate model for this data} \end{aligned}$$

We do not know λ, and so do not specify it in the model, so this falls into our second case.

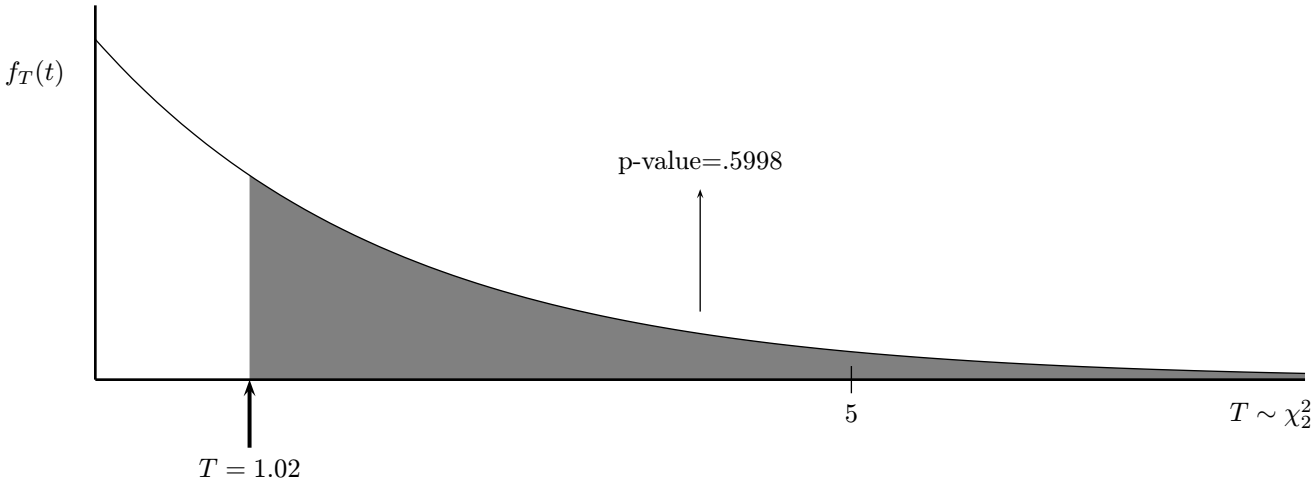
Let's create the necessary table of categories, observed counts, and expected counts all at once. We'll have the values 0, 1, and 2 represent their own categories, but group together 3 and above into a category (since together these will have somewhat low probability). Also, we need to calculate the MLE for λ before proceeding. We have seen that the MLE is $\bar{X}$, and so the estimate from this data is $\hat{\lambda} = 0.6941$. Putting all this together and making the appropriate expected probability calculations gives us

Values	Observed Prob	Observed Count	Expected Prob	Expected Count
0	.5059	43	.4995	42.46
1	.3176	27	.3467	29.47
2	.1529	13	.1203	10.23
3 and up	.0235	2	.0278	2.36
Total	1.00	85	1.00	85

Our test statistic calculation is

$$T = \frac{(43 - 42.46)^2}{42.46} + \frac{(27 - 29.47)^2}{29.47} + \frac{(13 - 10.23)^2}{10.23} + \frac{(2 - 2.36)^2}{2.36} = 1.02.$$

The statistic has a chi-squared distribution with (4-1-1)=2 degrees of freedom. So, the p-value is $\mathrm{P}(T > 1.02) = 1 - F_T(1.02) = 0.5998$, and our picture looks like



The p-value is quite large, suggesting that our data is very much in line with following a Poisson distribution. The closeness of the observed and expected counts makes this pretty clear. Our conclusion is that we do not have enough evidence to reject the null hypothesis, and so we do believe that a Poisson probability model does a good job describing this data.

2.11.4 More on Expected Counts

One of the rules of thumb we mentioned when discussing the general setup of the goodness-of-fit test was that the expected counts for any particular category should not be too low. Let's say more about this now.

This discussion will require a little more background on the test statistic and distribution used for this test. We have stated that the test statistic in Formula (2.13) has a chi-squared distribution if the null hypothesis is true, with degrees of freedom calculated depending on which of the two situations we were in.

But, this is only half of the story. Actually, this is the correct distribution ONLY IF the expected counts are not too low (and even in this case is just a close approximation of the correct distribution). This seems a little evasive, but there actually is no one perfect definition for what is meant by "not too low" in this case. A common rule of thumb that has been given by statisticians is that all expected counts should be greater than one, and most expected counts should be greater than five, with no more than 20% being less than five. We will go by this rule of thumb.

Looking back to our two example calculations, the rule of thumb was true in each case. Each time, we had all expected counts greater than one, and two of our ten categories (or 20%) had expected counts less than five, which just meets our definition. So it was in fact appropriate to use the chi-squared distribution to make our p-value calculations in those examples. However, in the Poisson example, one of our four categories (or 25%) had an expected count less than five, so we should have been aware of this fact at the time. See below for how we could have dealt with it.

If you do come across a situation where you have defined categories, but wind up not meeting this rule of thumb, there is a way to fix it. Simply, just redefine the categories by combining one or more together in a logical way. The categories we defined in the protein degradation example were really somewhat arbitrary. There are many ways to define them, some with more categories, some with fewer. We could have, for example, had just nine categories, with the ninth covering the range $[8, \infty)$. This would in essence combine the last two of our ten categories into one, and result in a category with higher expected probability and therefore a higher expected count. In the gamma distribution version of that example, the expected count for this ninth and final category would have been 10.73, combining two categories whose individual expected counts were 3.05 and 7.68. If we had trouble with having too low expected counts, this would have served to remove one of our "too low" categories.

For the Poisson model, we could have had just three categories by combining the last two. Then, the third category would have been all values 2 and greater, and this category would have had an expected count of 12.59. Our rule would have been satisfied. On your own, try re-doing that test after combining these categories.

To summarize, in most cases you should be able to create a categorization that meets our rule of having expected counts "not too low" by combining categories if necessary.

2.12 Likelihood Ratio Tests

Another set of very common testing methodologies are called *likelihood ratio tests*, or *LRTs*. We will hold off more thorough discussion of these types of tests until later in the course.

2.13 Tests for Individual Parameters

A very general class of tests that are often conducted are tests about a single parameter of a distribution. Although there are many ways to conduct such tests, and these ways can be very different depending on the distribution in question, we will only discuss a general method that works well across many different situations. In particular, it is an approximation to more correct methods that is a quite good approximation if the sample size is large.

Again, the advantage of this approximate technique is that one basic concept can be applied to almost any situation we come across. Otherwise, we would have to develop a special methodology for every distribution and parameter for which we wanted to conduct tests. The analogy is our discussion of confidence intervals for parameters; we only mentioned an approximate method that works well across many situations if the sample size is large.

To setup the notation, we have a random sample $X_1, \ldots, X_n$ with distribution $f_X(x)$ with unknown parameter θ. The distribution f can be either discrete or continuous. We'll use the general notation f to represent the probability distribution even for discrete random variables (we had used p_X previously) to keep the notation simple.

Our hypotheses have to do with the unknown θ, and will take one of the following three forms (along with related terminology):

$\mathrm{H}_0 : \theta = \theta_0$	$\mathrm{H}_0 : \theta = \theta_0$	$\mathrm{H}_0 : \theta = \theta_0$
$\mathrm{H}_a : \theta < \theta_0$	$\mathrm{H}_a : \theta > \theta_0$	$\mathrm{H}_a : \theta \neq \theta_0$
Left-tailed test	**Right-tailed test**	**Two-tailed test**

where θ_0 is a constant that represents some hypothesized value for θ.

The key thing is the alternative hypothesis. Our research question should be phrased along the lines of one of the three alternative hypotheses above. Are we trying to prove that the parameter is less than the specified constant, greater than it, or more generally, not equal to it? The choice of which alternative hypothesis type to use leads to us conducting what is called either a left-tailed, right-tailed, or two-tailed test as the table indicates.

The test statistic used to conduct such a test relies on the properties of maximum likelihood estimators that we have discussed. Namely, if $\hat{\theta}$ is the MLE of θ and we have a large sample size, we can say that $\hat{\theta}$ approximately has a normal distribution with mean θ (i.e., it is an approximately unbiased estimator), and standard deviation based on the typical calculation for that estimator.

We then conduct the test by using the following test statistic:

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$$

where any θ's in the formula are replaced by the hypothesized value θ_0 in the null hypothesis, and any other unknown parameters are replaced by their maximum likelihood estimates.

We refer to this statistic as Z for a good reason. The distribution it has if the null hypothesis is true is a standard normal distribution. We'll use this fact to look up the p-value for such a test.

With regard to the p-value calculation, these tests require us to put a little more thought into the definition of p-value. In particular, the part of the definition that refers to "more extreme" needs to be considered further. We mentioned earlier that these words will sometimes have a different interpretation for different tests.

If the test is a left-tailed test, then "more extreme" means "even further to the left" or "even lower." This is because if the alternative hypothesis is that the parameter is *less than* a certain value, our evidence is only strongly against the null if the test statistic is quite low, or in the left tail of the test statistic's distribution. In this case, large values of the test statistic wouldn't be considered evidence against H_0 , and in fact could only be considered as evidence favoring H_0 . So large values of the test statistic shouldn't be included when measuring the p-value in this case.

The inverse argument can be made if the alternative hypothesis is that the parameter is *greater than* the specified value, and so won't be elaborated on further.

However, let's discuss the two-tailed test situation in which the alternative hypothesis is that the parameter is *not equal to* the specified value. Here, both extremely small and large values of the test statistic would be considered as evidence against H_0 . So, our p-value calculation should take extreme values in both tails into account. We'll see the result of this more clearly in the example below.

Let's look at an example of testing for the value of a binomial proportion. Suppose $X_1, \ldots, X_n \sim \text{Binomial}(n, p)$. We can relate this to a study where we observe n random individuals and for each record whether or not they have a certain genetic disease. p is the percentage a research hypothesis regarding the parameter p. We would like to see if our data provides significant evidence to make the claim that $p > .04$. This may have some significance that will have an important impact on later research, or maybe is of important in and of itself. We won't believe this hypothesis unless we have strong evidence to believe it, and so it should be our alternative hypothesis. So our test is

$$H_0 : p = .04$$
$$H_a : p > .04$$

This falls into the class of tests we are discussing; it is a test regarding a single parameter of the probability model, and in fact is a right-tailed test.

Assuming that we have a large sample, the test procedure we discussed says that the test statistic we should use is

$$Z = \frac{\hat{p} - p}{\sigma_{\hat{p}}}$$

where unknown p's in the formula are replaced by the value p_0 which is .04 in this example (the hypothesized value). If we recall our initial discussion of estimation for the binomial distribution, we had calculated

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

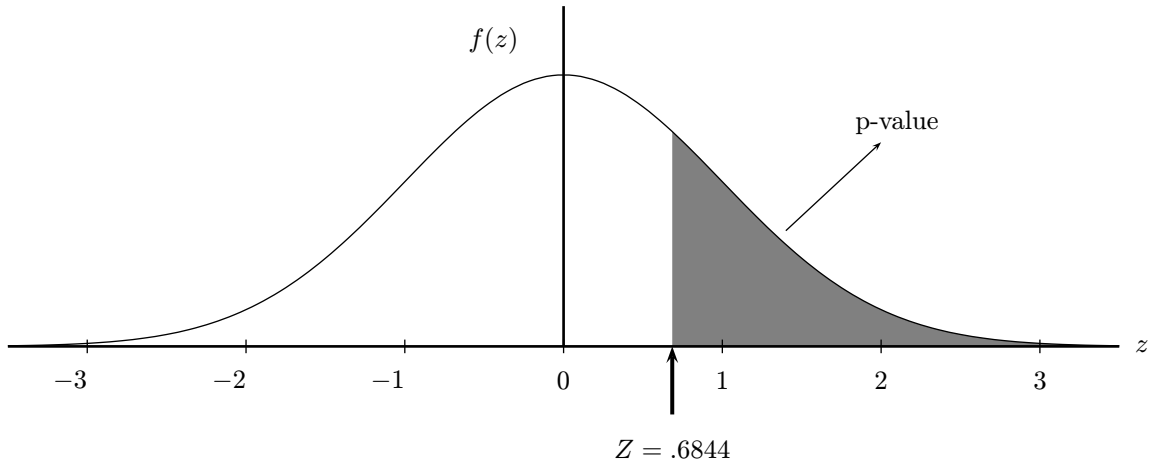
So our test statistic becomes

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\hat{p} - .04}{\sqrt{\frac{.04(1-.04)}{n}}}$$

Now, suppose we collect this data, and find that 7 in a sample of 136 have the disease. The value for $\hat{p}$ would be 7/136=.0515, and we can continue the above calculation

$$Z = \frac{\hat{p} - .04}{\sqrt{\frac{.04(1-.04)}{n}}} = \frac{.0515 - .04}{\sqrt{\frac{.04(1-.04)}{136}}} = 0.6844$$

This is the calculated value of our test statistic. To find the p-value, we must calculate the area pictured below:



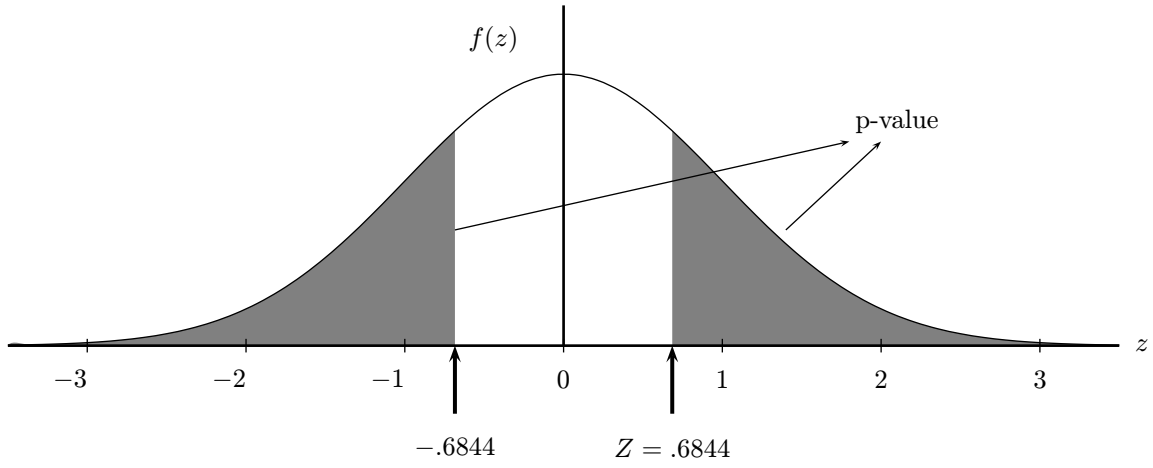
We consider only the values further to the right of the calculated test statistic as more extreme since this was a right tailed test. The calculation is p-value=P($Z > .6844$) $= 1 - F_Z(.6844) = 0.2469$. In Excel, we would compute

Excel Formula

```
1-NORMSDIST(.6844)=.2469
```

So, the p-value is somewhat large, leading us to not reject the null hypothesis. Even though the sample proportion was larger than the hypothesized value of .04, we do not have enough evidence to make the strong claim that the true proportion with the disease is greater than .04. The result is telling us that there is too large a chance that the true p is .04 or lower, and our data suggesting p is larger than .04 was an accidental result of random sampling. Note that it may not have been an accidental result, but we just don't have enough evidence to say otherwise at this point.

If the test above happened to have been two-tailed, with $H_a : p \neq .04$, everything would be the same except the p-value calculation. As mentioned above, we would need to consider values in both directions as extreme, and so the area we would need to calculate would be



The area includes values both to the right of +.6844 and to the left of -.6844. To make this a little easier, we should notice that the Z-curve is symmetric around 0, and so these two separate areas are exactly the same. So, we really just need to calculate one of the areas (which we have already done) and double it to get the total p-value. The calculation is p-value=2P($Z > .6844$) = 0.4937. So, if this had been a two-tailed test, the evidence against H_0 would have been even weaker.

Chapter 3

Genetic Frequencies and Hardy-Weinberg Equilibrium

A very common and important problem in statistical genetics is the estimation of genotype and/or allele frequencies, as well as phenotype frequencies. The applications of these estimation problems are wide-ranging, from locating disease genes to forensic DNA testing.

3.1 Notation

Let us begin by setting up the notation we will use when discussing genetic frequencies. First, consider a single gene which we'll refer to as **A**. We will refer to the number of different allelic forms that exist in the population by the symbol k. We refer to these different allelic forms by the notation A_1, ..., A_k (as seen in Chapter 2). For example, $k = 3$ for the ABO blood group gene. The general notation A_1, A_2, and A_3 refer to the more specific allele references **A**, **B**, and **O** for this gene.

Now, for our generic gene **A**, consider a population of N individuals. Since each individual has two alleles for the gene (unless it is a sex-linked gene), there are $2N$ total alleles to account for in this population. In theory, we could determine how many A_1 alleles existed, how many A_2 alleles existed, and so on for all k allelic forms. By taking these allele counts and dividing by $2N$ for each, we would arrive at the *population allele frequencies*. We will refer to them as $P_{A_1}, P_{A_2}, \ldots, P_{A_k}$. Of course, it must be that

$$\sum_{i=1}^{k} P_{A_i} = 1.$$

It is important to realize that these population allele frequencies are population parameters. Because it is not realistic to think that we count alleles across an entire population, they are unknown, but real, quantities. One problem we would like to solve, for any particular gene, is to determine good estimates of these population allele frequencies.

If we now concentrate on genotypes instead of alleles, we can introduce the concept of *population genotype frequencies*. For our gene **A** with k possible allelic forms, there are $k(k+1)/2$ possible different genotypes in the population. By the same argument as above, in theory, for the N individuals in the population, we could determine how many A_1A_1 genotypes there were, how many A_1A_2 there were, and so on for all the genotypes. Dividing these counts by N would give us the population genotype frequencies. We will use the generic notation $P_{A_iA_j}$ to refer to the population genotype frequency for genotype A_iA_j, where $1 \leq i \leq j \leq k$. Of course, these frequencies must follow the constraint of summing to one:

$$\sum_{i=1}^{k}\sum_{j=i}^{k} P_{A_iA_j} = 1.$$

As with population allele frequencies, we consider population genotype frequencies as unknown parameters because of the enormous effort it would take to determine them exactly.

Also, notice that population allele frequencies can be calculated if given the population genotype frequencies. For any allele i, its population frequency can be written as

$$P_{A_i} = 2P_{A_iA_i} + \sum_{i \neq j} P_{A_iA_j}$$

Finally, we can follow the same logic to introduce *population phenotype frequencies*, or any other population-based frequencies for that matter. The standard notation we will use throughout for such population frequencies is the symbol P subscripted by a description of the category of individual the frequency represents. An example of notation for population phenotype frequencies would be P_A, P_B, P_{AB}, and P_O to represent the population frequencies for the ABO blood group locus.

Keep in mind that the term population need not be equated with "all people" or similar statements for other species. We may be studying, for example, the population of all African-Americans living in the southeastern United States, or all men with heart disease, or all tomato plants not resistant to a certain insect.

In the discussion above, we have alluded to the fact that one problem we will need to solve is to find reasonable methods to estimate these unknown population frequencies, based only on a (usually small) sample of the population. Generally, we will use techniques of maximum likelihood estimation, as discussed in Chapter 2. There are other issues we will want to investigate regarding such frequencies. In particular, for allele frequencies at a gene, we would like to investigate how they relate to each other and to the genotype frequencies for the gene. The most important such relationship is Hardy-Weinberg equilibrium.

3.2 Hardy-Weinberg Equilibrium

There is often an important relationship between genotype and allele frequencies that is present in populations. This relationship is called *Hardy-Weinberg Equilibrium* or *HWE*. First, we will discuss the genetic

assumptions that lead a population to be in HWE, then discuss what it means from a mathematical perspective.

Consider any population of individuals. It is said that the population is in Hardy-Weinberg equilibrium is it satisfies the following seven assumptions (Lange, 2003):

1. The population size is infinite;
2. The generations are discrete (all individuals in a generation are born at the same time);
3. Matings between individuals occur completely at random;
4. There are no natural selection pressures on the population;
5. There is no migration into or out of the population by any individuals;
6. No genetic mutations occur;
7. The two sexes have equal genotype frequencies to start.

Taken together, we will often refer to this collection of assumptions as the *Hardy-Weinberg Assumptions*, or sometimes as just *random mating*. We will not discuss the details underlying each of these assumptions as they are each chapter-worthy topics in their own right. There are many helpful texts that can be used to study these further. Even without an in-depth understanding of these assumptions, it should be clear that they can never be completely true for any population (the first assumption of infinite population size is enough of a problem itself). However, in the discussion that follows, we will consider the assumptions to be close enough to true that the implications to be discussed will still be valid.

If the assumptions are true (or nearly true), a special relationship between genotype frequencies and allele frequencies at the locus will be true. Suppose we have a locus with k alleles $A_1, \cdots, A_k$ having population frequencies $P_{A_1}, \cdots, P_{A_k}$. The $k(k+1)/2$ possible genotypes have population frequencies given by $P_{A_iA_j}$, $i \leq j$.

The HWE law that relates the genotype frequencies to allele frequencies is as follows:

$$\begin{aligned} P_{A_iA_i} &= P_{A_i}^2, \quad i = 1, \ldots, k \text{ (for homozygotes)} \\ P_{A_iA_j} &= 2P_{A_i}P_{A_j}, \quad j > i = 1, \ldots, k \text{ (for heterozygotes)} \end{aligned} \tag{3.1}$$

That is, any homozygote genotype frequency is the square of the relevant allele frequency, and a heterozygote frequency is twice the product of the relevant allele frequencies. If the relationship between genotype and allele frequencies at a locus have this relationship, we will say that the locus is in Hardy-Weinberg Equilibrium, or that the genotype and allele frequencies is in HWE. The frequencies themselves will be called Hardy-Weinberg proportions.

The Hardy-Weinberg proportions, as seen in Equation (3.1), are actually a set of $k(k+1)/2$ equations (do not be fooled by the fact that there are just two lines in the equation). If there are only two alleles at the locus ($k = 2$), then HWE reduces to the following three explicit equations:

$$\begin{aligned} P_{A_1A_1} &= P_{A_1}^2 \\ P_{A_1A_2} &= 2P_{A_1}P_{A_2} \\ P_{A_2A_2} &= P_{A_2}^2 \end{aligned} \tag{3.2}$$

3.3 Derivation of Hardy-Weinberg Proportions

It is instructive, albeit lengthy, to see how this law comes about if we truly have random mating occurring in a population. Let's do it for a locus with two alleles which we'll call A and a for ease of notation. Let's start with a population that has genotype frequencies in the first generation as $P_{AA}^{(0)}, P_{Aa}^{(0)}$, and $P_{aa}^{(0)}$. The (0) superscript makes it clear that these are the starting frequencies. Notice that, by definition, the starting allele frequencies are $P_A^{(0)} = P_{AA}^{(0)} + \frac{1}{2}P_{Aa}^{(0)}$ and $P_a^{(0)} = P_{aa}^{(0)} + \frac{1}{2}P_{Aa}^{(0)}$.

In constructing the next generation (i.e., generation (1)), we consider that random mating occurs in this population. That is, individuals mate with each other regardless of their genotypes, and all the other assumptions of HWE are also true. We can envision the result using a couple of tables. The first gives the possible matings and their probabilities.

	AA	Aa	aa
AA	${P_{AA}^{(0)}}^2$	$2P_{AA}^{(0)}P_{Aa}^{(0)}$	$2P_{AA}^{(0)}P_{aa}^{(0)}$
Aa		${P_{Aa}^{(0)}}^2$	$2P_{Aa}^{(0)}P_{aa}^{(0)}$
Aa			${P_{aa}^{(0)}}^2$

Since the matings are symmetric, the lower left side is left blank, and combined with the relevant upper right side cells using the 2 multiplier. Now, the relevant offspring proportions are given by:

	AA	Aa	aa
AA	All AA	$\frac{1}{2}AA$, $\frac{1}{2}Aa$	All Aa
Aa		$\frac{1}{4}AA$, $\frac{1}{2}Aa$, $\frac{1}{4}aa$	$\frac{1}{2}Aa$, $\frac{1}{2}aa$
Aa			All aa

Let's use these to calculate the expected proportion of AA genotypes in the next generation. We'll call this value $P_{AA}^{(1)}$. We will calculate this value using a combination of the two tables above. For example, since all matings of the type $AA \times AA$ will result in all offspring having genotype AA (from the second table), and this mating happens with proportion ${P_{AA}^{(0)}}^2$ of all matings (from the first table), then the first term below will be ${P_{AA}^{(0)}}^2 \times 1$. Other terms will come by looking for the other situations where AA offspring will occur, in what proportions, and from what mating. Putting all of it together, we will get

$$\begin{aligned} P_{AA}^{(1)} &= {P_{AA}^{(0)}}^2 \times 1 + 2P_{AA}^{(0)}P_{Aa}^{(0)} \times \frac{1}{2} + {P_{Aa}^{(0)}}^2 \times \frac{1}{4} \\ &= {P_{AA}^{(0)}}^2 + P_{AA}^{(0)}P_{Aa}^{(0)} + \frac{1}{4}{P_{Aa}^{(0)}}^2 \end{aligned}$$

A similar argument to calculate $P_{Aa}^{(1)}$ gives us

$$P_{Aa}^{(1)} = P_{AA}^{(0)}P_{Aa}^{(0)} + 2P_{AA}^{(0)}P_{aa}^{(0)} + \frac{1}{2}{P_{Aa}^{(0)}}^2 + P_{Aa}^{(0)}P_{aa}^{(0)}.$$

We could also calculate $P_{aa}^{(1)}$, but for now let's focus on the above two quantities. From these, we can directly calculate the frequency of allele A in this new generation as

$$\begin{aligned}
P_A^{(1)} &= P_{AA}^{(1)} + \frac{1}{2}P_{Aa}^{(1)} \\
&= \left({P_{AA}^{(0)}}^2 + P_{AA}^{(0)}P_{Aa}^{(0)} + \frac{1}{4}{P_{Aa}^{(0)}}^2\right) \\
&\quad + \frac{1}{2}\left(P_{AA}^{(0)}P_{Aa}^{(0)} + 2P_{AA}^{(0)}P_{aa}^{(0)} + \frac{1}{2}{P_{Aa}^{(0)}}^2 + P_{Aa}^{(0)}P_{aa}^{(0)}\right) \\
&= \left({P_A^{(0)}}^2\right) + \frac{1}{2}\left(2P_A^{(0)}P_a^{(0)}\right) \\
&= {P_A^{(0)}}^2 + P_A^{(0)}P_a^{(0)} \\
&= P_A^{(0)}\left(P_A^{(0)} + P_a^{(0)}\right) \\
&= P_A^{(0)}
\end{aligned}$$

So, after a generation of random mating, the expected allele frequency for allele A has not changed. Of course, since they add to 1, it must also be true that the frequency of allele a has not changed either, and so $P_a^{(1)} = P_a^{(0)}$.

This is interesting, but does not in itself show that HWE will continue to hold in generation (1). We now need to revisit the formulas for $P_{AA}^{(1)}$ and $P_{Aa}^{(1)}$. First, from above we saw that

$$P_{AA}^{(1)} = {P_{AA}^{(0)}}^2 + P_{AA}^{(0)}P_{Aa}^{(0)} + \frac{1}{4}{P_{Aa}^{(0)}}^2.$$

We notice that the right side can be written as ${P_A^{(0)}}^2$, which in turn we just found to be equal to ${P_A^{(1)}}^2$. So, combining these, we now have

$$P_{AA}^{(1)} = {P_A^{(1)}}^2$$

which shows that HWE holds at least for the AA homozygote frequency. But using the same method shows that HWE holds for all three genotypes. For example, for the heterozygote genotype

$$\begin{aligned}
P_{Aa}^{(1)} &= P_{AA}^{(0)}P_{Aa}^{(0)} + 2P_{AA}^{(0)}P_{aa}^{(0)} + \frac{1}{2}{P_{Aa}^{(0)}}^2 + P_{Aa}^{(0)}P_{aa}^{(0)} \\
&= 2P_A^{(0)}P_a^{(0)} \\
&= 2P_A^{(1)}P_a^{(1)}.
\end{aligned}$$

So the heterozygote genotype probability is twice the product of the two allele frequencies. Finally, the correct HWE relationship for the other homozygote holds as well: $P_{aa}^{(1)} = {P_a^{(1)}}^2$.

Together, all of this shows that regardless of the relationship between genotypic and allele frequencies for a locus at a particular point in time (remember, we started our derivation with generic frequencies), once a generation of random mating occurs and all the other HWE assumptions are true, the genotype and allele frequencies will be in HWE in that next generation.

Of course, exact patterns of random mating never really occur in real populations, and additionally all of the other HWE assumptions will never be exactly true. But, it has been found that together they are close enough to true (or possibly display small departures that tend to cancel each other) that many loci are found to be at least nearly in Hardy-Weinberg equilibrium. Our main concern will be to determine if the genotype and allele frequencies are in HWE. We will not worry much about whether all of the genetic assumptions leading to HWE are exactly true or not.

Some problems we will investigate further later in the chapter are to estimate allele frequencies if we can assume that the locus is in HWE, and to determine whether or not a locus can be said to be in HWE.

3.4 Maximum Likelihood Estimation of Frequencies

We will now discuss genetic frequencies from a statistical point of view. We want to estimate such population frequencies from collected data using maximum likelihood methods. Statistically, the setup of the problem is fairly straightforward. For example, if we have a random sample of genotypes for a population, we can regard this as sampling from a multinomial distribution since each individual has one of a certain number of possible genotypes, and each individual is independent in the random sample. The same argument applies for a sample of alleles. So, we will first begin with a more detailed look at the multinomial distribution and resulting maximum likelihood estimates.

Recall the setup for a multinomial distribution. We have $X_1, \ldots, X_k \sim \text{Multinomial}(n, p_1, \ldots, p_k)$, where $\sum_{i=1}^k X_i = n$ and $\sum_{i=1}^k p_i = 1$, and $\text{p}_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \frac{n!}{\Pi x_i!}\ p_1^{x_1}\ \cdots\ p_k^{x_k}$. The parameters $p_1, \ldots, p_k$ are unknown, and are to be estimated. Actually, we need to be a little careful and realize that there are actually only $k-1$ parameters, since we can write p_k as a function of the others: namely, $p_k = 1 - p_1 - \cdots - p_{k-1}$.

Let's find the maximum likelihood estimators for $p_1, \ldots, p_{k-1}$. The situation is a little more involved from a calculus standpoint than we had come across before. Since there are $k-1$ parameters to estimate, the likelihood function is a function of $k-1$ variables, not just 1, and so there are $k-1$ variables that we have to take derivatives with respect to. We'll start by writing the joint pmf in the notation of a likelihood function:

$$\text{L}(p_1, \ldots, p_{k-1}) = \frac{n!}{\Pi x_i!}\ p_1^{x_1}\ \cdots\ p_{k-1}^{x_{k-1}}\ (1 - p_1 - \cdots - p_{k-1})^{x_k}.$$

Notice that we have written the p_k term in terms of the other p's. Now the log-likelihood is:

$$\ln \text{L}(p_1, \ldots, p_{k-1}) = \ln n! - \ln(\Pi x_i!) + x_1 \ln p_1 + \cdots + x_{k-1} \ln p_{k-1} + x_k \ln(1 - p_1 - \cdots - p_{k-1}).$$

Now we can start taking derivatives:

$$\begin{aligned}\frac{\partial \ln \mathrm{L}}{\partial p_1} &= \frac{x_1}{p_1} - \frac{x_k}{(1-p_1-\cdots-p_{k-1})} = 0 \\ &\vdots \\ \frac{\partial \ln \mathrm{L}}{\partial p_{k-1}} &= \frac{x_{k-1}}{p_{k-1}} - \frac{x_k}{(1-p_1-\cdots-p_{k-1})} = 0\end{aligned}$$

Solving these $k-1$ equations requires a little algebra, but it is much easier to use the Lagrange multiplier technique to maximize this likelihood instead of the standard method above. This technique can be used when there is a constraint to the parameters. In this case, the constraint is that $1-\sum p_i = 0$. This constraint is multiplied by λ and added to the log-likelihood before taking derivatives. This also allows us to keep the last parameter p_k in the equation as is instead of rewriting it. With this in place, the log-likelihood now becomes:

$$\ln \mathrm{L}(p_1, \ldots, p_k, \lambda) = C + \sum_{i=1}^{k} x_i \ln p_i + \lambda(1 - \sum p_i).$$

Differentiating with respect to each p_i and setting equal to 0 gives:

$$\begin{aligned}\frac{\partial \ln \mathrm{L}}{\partial p_1} &= \frac{x_1}{p_1} - \lambda = 0 \\ &\vdots \\ \frac{\partial \ln \mathrm{L}}{\partial p_k} &= \frac{x_k}{p_k} - \lambda = 0\end{aligned}$$

To solve these k equations, we can add them up to get

$$\sum_{i=1}^{k} x_i = \lambda \sum_{i=1}^{k} p_i.$$

The left side above is n and the right side is λ. So we have $\lambda = n$, which we can plug back into each of the n equations. Doing this for each equation gives the estimates:

$$\begin{aligned}\hat{p_1} &= \frac{x_1}{n} \\ &\vdots \\ \hat{p_k} &= \frac{x_k}{n}\end{aligned}$$

So, the maximum likelihood estimates for the probability parameters are just the sample proportions of individuals that fell into that category, which makes logical sense.

Let's take a closer look at any particular such estimate $\hat{p_i}$. Its mean is

$$\mu_{\hat{p_i}} = \frac{1}{n}\mu_{x_i} = \frac{1}{n}\ (np_i) = p_i,$$

using our knowledge of the multinomial distribution means. This shows that it is an unbiased estimator for p_i. Its variance can be calculate as:

$$\sigma^2_{\hat{p_i}} = \frac{1}{n^2}\sigma^2_{x_i} = \frac{1}{n^2}\ [np_i(1-p_i)] = \frac{p_i(1-p_i)}{n}.$$

So we can calculate individual approximate 95% confidence intervals for p_i by the formula:

$$\hat{p_i} \pm 2 \times \sqrt{\frac{\hat{p_i}(1-\hat{p_i})}{n}}$$

3.4.1 Multinomial Genetic Data

Now let's look at the problem specifically from a genetic point of view. Suppose a locus with two alleles A and a is in HWE. We will collect genotypic data and estimate the two allele frequencies P_A and P_a. Since the locus is in HWE, we can write the genotype frequencies in terms of the allele frequencies as

$$\begin{aligned} P_{AA} &= P_A^2 \\ P_{Aa} &= 2P_AP_a = 2P_A(1-P_A) \\ P_{aa} &= P_a^2 = (1-P_A)^2 \end{aligned}$$

where we have written the frequency P_a as $1 - P_A$ since they must add to 1.

Now, we have a multinomial model with three categories (i.e., the three genotypes), but only one unknown parameter, namely P_A. Let's use the MLE method to find the maximum likelihood estimator for P_A.

Our model is $(X_{AA}, X_{Aa}, X_{aa}) \sim \text{Multinomial}(n, P_A^2, 2P_A(1-P_A), (1-P_A)^2)$. This is a good example of a situation where the model parameters are related to each other. Here, all three multinomial probabilities can be written in terms of just the one parameter P_A. This has no bearing on our ultimate goal in maximum likelihood estimation, which is to write down the log-likelihood of the parameter, take derivatives, and solve.

Now, since this is a multinomial model, the likelihood function for P_A (or in other words, the joint pmf for X_{AA}, X_{Aa}, and X_{aa}) is

$$\text{L}(P_A) = \frac{n!}{x_{AA}!x_{Aa}!x_{aa}!}\ \left(P_A^2\right)^{x_{AA}}\ \left[2P_A(1-P_A)\right]^{x_{Aa}}\ \left[(1-P_A)^2\right]^{x_{aa}}$$

and the log-likelihood is

$$\begin{aligned}\ln \mathrm{L}(P_A) &= C + 2x_{AA}\ln P_A + x_{Aa}\ln\big[2P_A(1-P_A)\big] + 2x_{aa}\ln\big(1-P_A\big) \\ &= C + (2x_{AA}+x_{Aa})\ln P_A + (x_{Aa}+2x_{aa})\ln(1-P_A).\end{aligned}$$

Taking the derivative of this with respect to P_A gives:

$$\frac{\partial \ln \mathrm{L}}{\partial P_A} = \frac{2x_{AA}+x_{Aa}}{P_A} - \frac{x_{Aa}+2x_{aa}}{1-P_A}$$

Setting this equal to 0 and solving for P_A gives us the MLE

$$\hat{P_A} = \frac{1}{2n}(2x_{AA}+x_{Aa})$$

which just counts up the number of A alleles in the sample and divides by the total number of alleles $2n$.

Example 3.1: **Genotypic data:** Suppose we have genotypic data available at a locus with two alleles A and a. The three possible genotypes are, of course, AA, Aa, and aa, to which we will assign unknown population probabilities p_{AA}, p_{Aa}, and p_{aa}. With a random sample of size n, we will let X_{AA}, X_{Aa}, and X_{aa} represent the observed counts for the three genotypes. Because we are randomly sampling individuals, our probability model is a multinomial distribution:

$$(X_{AA}, X_{Aa}, X_{aa}) \sim \text{Multinomial}(n, p_{AA}, p_{Aa}, p_{aa})$$

Suppose we collect data from a random sample of 104 individuals, resulting in 32 AA genotypes, 60 Aa genotypes, and 12 aa genotypes. Our genotype frequency estimates will be:

$$\begin{aligned}\widehat{p_{AA}} &= \frac{32}{104} = .3077 \\ \widehat{p_{Aa}} &= \frac{60}{104} = .5769 \\ \widehat{p_{aa}} &= \frac{12}{104} = .1154\end{aligned}$$

A 95% confidence interval estimate for p_{AA} is:

$$\begin{aligned}&\widehat{p_{AA}} \pm 2 \times \sqrt{\frac{\widehat{p_{AA}}(1-\widehat{p_{AA}})}{n}} \\ =&.3077 \pm 2 \times \sqrt{\frac{.3077 \times (1-.3077)}{104}} \\ =&.3077 \pm .0905 \\ =&(.2172, .3982)\end{aligned}$$

Similar CI estimates can be calculated for the other genotypic frequencies. Notice that this is a somewhat large range in our CI estimate. This can be decreased to come up with a more precise estimate by increasing the sample size (which makes the denominator on the right side larger, and so the entire right side smaller).

■

Notice that if we do have genotypic data, we can use the same data and same procedure to estimate allele frequencies as well as genotype frequencies. Our random sample of n individuals can also be viewed as a random sample of $2n$ alleles. So the allele counts will have a multinomial distribution with sample size $2n$. We can write our probability model as:

$$(X_A, X_a) \sim \text{Multinomial}(2n, p_A, p_a).$$

Of course, with only two alleles at the locus, this really reduces to being a Binomial distribution, but we can think of it either way since the Binomial is just a special case of the Multinomial. So our maximum likelihood estimates of the unknown allele frequencies p_A and p_a are:

$$\hat{p_A} = \frac{X_A}{2n} \quad \text{and} \quad \hat{p_a} = \frac{X_a}{2n}$$

Example 3.2: **Genotypic data (cont.):** If we continue the example above, we notice that there were a total of $2X_{AA} + X_{Aa} = 2(32) + 60 = 124$ A alleles in the sample. This is X_A. So our estimate of p_A is $\hat{p_A} = \frac{124}{2*104} = .5962$. Similarly, $X_a = 2(12) + 60 = 84$ and so $\hat{p_a} = \frac{84}{2*104} = .4038$. An approximate 95% confidence interval for p_A would be

$$\begin{aligned}
&\hat{p_A} \pm 2 \times \sqrt{\frac{\hat{p_A}(1-\hat{p_A})}{2n}} \\
=&.5962 \pm 2 \times \sqrt{\frac{.5962 \times (1 - .5962)}{208}} \\
=&.5962 \pm .0680 \\
=&(.5282, .6642)
\end{aligned}$$

■

The method for computing allele frequency estimates in the above example is often called *gene counting* since all we have done is count up the number of each allele in the sample of genotype data to come up with our estimate. However, if it is not possible to have genotypic data, then it might be impossible to use this method to estimate allele frequencies. Often only phenotypic data is available for a gene. For example, if the allele A is dominant to a and only phenotypic data was collected (for the two possible phenotypes), then we wouldn't know how many of the dominant phenotype actually had genotype AA and how many had Aa, and so couldn't do this gene counting, and so couldn't directly use the multinomial model to make frequency estimates. We will soon discuss another method for doing this.

3.4.2 Non-HWE Locus MLEs

What gets a little more interesting is when we consider a two-allele locus which is not in HWE. A common way to model the genotype frequencies that allows us to measure the amount of departure from HWE is

$$\begin{aligned} P_{AA} &= P_A^2 + P_A(1-P_A)f \\ P_{Aa} &= 2P_A(1-P_A)(1-f) \\ P_{aa} &= (1-P_A)^2 + P_A(1-P_A)f. \end{aligned}$$

The f parameter is a measure of how far from HWE the locus is. Notice that if $f = 0$, the proportions above are just the HW proportions. Any other value for f gives genotype frequencies that are different from the frequencies that would be expected under HWE.

Now in our multinomial model, we again have two unknown parameters, P_A and f. Our model is

$$\begin{aligned} (X_{AA}, X_{Aa}, X_{aa}) \sim \text{Multinomial}\big[n,\ & P_A^2 + P_A(1-P_A)f, \\ & 2P_A(1-P_A)(1-f), \\ & (1-P_A)^2 + P_A(1-P_A)f\big]. \end{aligned}$$

and now our likelihood and log-likelihood functions are

$$\begin{aligned} \mathrm{L}(P_A, f) = &\frac{n!}{x_{AA}!x_{Aa}!x_{aa}!}\ \big[P_A^2 + P_A(1-P_A)f\big]^{x_{AA}}\ \big[2P_A(1-P_A)(1-f)\big]^{x_{Aa}} \\ &\times \big[(1-P_A)^2 + P_A(1-P_A)f\big]^{x_{aa}} \end{aligned}$$

and the log-likelihood is, after some simplification,

$$\begin{aligned} \ln \mathrm{L}(P_A, f) = C &+ (x_{AA} + x_{Aa}) \ln P_A + x_{AA} \ln\big[P_A + (1-P_A)f\big] \\ &+ (x_{Aa} + x_{aa}) \ln(1-P_A) + x_{Aa} \ln(1-f) \\ &+ x_{aa} \ln\big[(1-P_A) + P_A f\big]. \end{aligned}$$

Differentiating with respect to P_A and f gives

$$\begin{aligned} \frac{\partial \ln \mathrm{L}}{\partial P_A} = &\frac{x_{AA} + x_{Aa}}{P_A} - \frac{x_{Aa} + x_{aa}}{1-P_A} \\ &+ \frac{x_{AA}(1-f)}{P_A + (1-P_A)f} - \frac{x_{aa}(1-f)}{(1-P_A) + P_A f} \\ \frac{\partial \ln \mathrm{L}}{\partial f} = &\frac{x_{AA}(1-P_A)}{P_A + (1-P_A)f} - \frac{x_{Aa}}{1-f} + \frac{x_{aa}P_A}{(1-P_A) + P_A f} \end{aligned}$$

Now we would need to set these equal to zero and solve this system of two equations for the two unknowns P_A and f. Some algebraic manipulation, preferably using a computer package, finds the solution

$$\hat{f} = \frac{4x_{AA}x_{aa} - x_{Aa}^2}{(2x_{AA} + x_{Aa})(x_{Aa} + 2x_{aa})}$$
$$\hat{P_A} = \frac{1}{2n}(2x_{AA} + x_{Aa})$$

We see that the estimator for P_A (and therefore also the related estimator for P_a) is just the gene counting estimate from before. The estimator for f, however, is a little more difficult to understand naturally. A little bit of understanding can be gleaned from the numerator: if there are too many heterozygotes compared to expected under HWE (i.e., if the genotypic counts are such that $x_{Aa}^2 > 4x_{AA}x_{aa}$, then the estimate of f will be negative, otherwise it will be positive. The denominator is just the product of allele counts.

Computing formulas for the variances of these estimators is a little more involved than previous variances we have calculated. This is because the formulas for the estimators themselves are much more complicated (particularly the formula for $\hat{f}$), involving products and ratios of random variables which lead to calculation difficulties.

However, there are methods for deriving an approximate formula for the variance of such estimators. One in particular that works well is called Fisher's Method or sometimes the Delta Method. It is approximate in that it leaves out terms divided by n^2 or higher powers of n. See Weir (1996) for more details. Here, for now at least, we will just write out these complicated variances without worrying about their derivation. The result is (again, see the Weir reference for details):

$$\sigma_{\hat{f}}^2 = \frac{1}{n}(1-f)^2(1-2f) + \frac{f(1-f)(2-f)}{2nP_A(1-P_A)}$$

and

$$\sigma_{\hat{P_A}}^2 = \frac{1}{2n}(1+f)P_A(1-P_A)$$

Of course, these formulas involve the unknown parameters f and P_A, so as before, we would substitute those with their estimates to make the final calculation.

Note that the variance of $\hat{P_A}$ is slightly more involved than we might have expected (and more involved than we have seen before) because the multinomial probability model for the genotypes itself is more involved. In the case where $f = 0$ (i.e., HWE holds), then it reduces to the same variance formula we have seen before.

Example 3.3: Another blood group gene in humans is the MN blood group. See http://www.mun.ca/biology/scarr/MN_bloodgroup.htm for more details and differences among various human populations. This blood group is helpful to study because there are two alleles (labeled M and N) and they are codominant, meaning that observing only phenotypes still allows us to know the genotypes exactly. In other words, a persons MN-blood type is either M, N, or MN which correspond to genotypes MM, NN, and MN.

Let's look at data presented by Crow (1986). In a sample of 208 from the human population of Bedouins, the following phenotype counts were found:

Phenotype	Genotype	Count
M	MM	119
MN	MN	76
N	NN	13

Let's use the multinomial model presented above that includes the parameter f, and estimate P_M, P_N and f for this population. Plugging into the maximum likelihood estimators above, we get:

$$\hat{P_M} = 0.755$$

$$\hat{P_N} = 1 - \hat{P_M} = 1 - 0.755 = 0.245$$

and

$$\hat{f} = \frac{4(119)(13) - (76)^2}{[2(119) + 76][2(13) + 76]} = .0129.$$

The estimates of P_M and P_N are just the gene counting estimates. The estimate of f is quite close to 0, suggesting that the genotype frequencies are not too far from being in HW proportions. Being slightly positive suggests that there is a small excess of homozygotes compared to what would be expected.

Let's also attach variances to these estimates and compute approximate confidence intervals. Using the variance formulas above, we have

$$\sigma^2_{\hat{P_M}} = \frac{1}{416}(1 + .0129)(.755)(.245) = .00045$$

giving an approximate 95% CI for P_M as

$$.755 \pm 2\sqrt{.00045} = .755 \pm .042 = (.713, 797).$$

For f we have

$$\sigma^2_{\hat{f}} = \frac{1}{208}(1 - .0129)^2(1 - 2 \times .0129) + \frac{.0129(1 - .0129)(2 - .0129)}{(416)(.755)(.245)} = .0049$$

with the approximate 95% CI

$$.0129 \pm 2\sqrt{.0049} = .0129 \pm .1399 = (-0.127, 0.1528).$$

■

3.4.3 The E-M Algorithm for Estimating Allele Frequencies

We can now return to our problem of estimating allele frequencies when we have phenotypic data and not genotypic data. We will consider the problem where there are two alleles A and a at a locus, and that A is dominant to a. We only have access to phenotypic data (such as disease status for a disease gene). This means that we will be able to count the total number of individuals $x_{AA} + x_{Aa}$, which will be the number with the dominant phenotype, but not break it down further into those two genotypes. We will also know the count x_{aa} which is the number with the recessive phenotype.

As seen before, simple gene counting techniques will not work because of the hidden genotypic data for the AA and Aa genotypes. We can think of this situation as being one where there is missing data. We don't know those two counts individually, but we do know their total. An algorithm called the E-M algorithm (for Expectation-Maximization) is suited to handle such situations. It is a very generic iterative technique, not at all specific to genetics, that can be used to find maximum likelihood estimates in this missing data type of situation. We will only discuss the algorithm as it applies to our allele frequency estimation problem.

First, one assumption we must make to proceed is that the locus is in HWE. So Equation (3.2) will be assumed true (or Equation (3.1) for any number of alleles). We will see why we need this to be true below.

The basic concept of the E-M algorithm is as follows. Recall that our main goal here is to come up with an estimate for the unknown allele frequencies. Our ensuing example will be for two alleles, so we have just one unknown frequency which we have called P_A (the other being $1 - P_A$).

1. Make a reasonable initial guess at the unknown frequency or frequencies.
2. Determine a reasonable way to improve this guess, based on the probability model and the HWE assumption.
3. Return to Step 2 over and over until the change in estimates from one step to the next is very small.

The mathematical concept described in Step 3 in the E-M algorithm is called *convergence*. Basically, this means that an iterative algorithm has finished doing its work when not much changes from one iteration to the next. The extreme case would be that there is precisely no change from one iteration to the next, in which case the algorithm will stay in that state forever (and so you definitely should stop). The definition of "very small change" depends on the situation. In the case of allele frequencies, a change on the order of 10^{-4} would certainly be considered quite small. This value is often called the *convergence criterion*.

Now let's return to the algorithm itself and how it will work for our allele frequency estimation problem. First, let's focus on estimating P_a by this algorithm, and our estimate of P_A will just be $1 - P_a$. Our first step is to come up with an initial guess for an estimate of P_a. There are many ways to do this. One quick way that comes to mind is that we could estimate the number of a alleles to be $2x_{aa}$ plus one-quarter of the total $2(x_{AA} + x_{Aa})$. Then our initial guess at P_a (which we'll superscript with (0)) would be

$$\hat{P}_a^{(0)} = \frac{2x_{aa} + \frac{1}{2}(x_{AA} + x_{Aa})}{2n}$$

Now, we need a way to improve this estimate. Notice that if we did have full genotypic data, we could write our improved estimate (which we'll superscript with (1)), or any estimate for that matter, as

$$\hat{P_a}^{(1)} = \frac{1}{2n}(x_{Aa} + 2x_{aa}) \tag{3.3}$$

but we are stuck not knowing x_{Aa}. However, using our HWE assumption, we know that the expected proportion of the total count $x_{AA} + x_{Aa}$ that can be attributed to the Aa genotype is

$$\widehat{P_{Aa}} = \frac{2\hat{P_a}^{(0)}(1 - \hat{P_a}^{(0)})}{(1 - \hat{P_a}^{(0)})^2 + 2\hat{P_a}^{(0)}(1 - \hat{P_a}^{(0)})}$$

which can be re-written as

$$\widehat{P_{Aa}} = \frac{2\hat{P_a}^{(0)}(1 - \hat{P_a}^{(0)})}{1 - (\hat{P_a}^{(0)})^2}$$

since the denominator was the sum of the HWE proportions for the first two genotypes, which is just 1 minus the HWE proportion for the third genotype. The formula on the right side is based on the initial estimate we made for P_a. Therefore, our estimate of the total number of Aa genotypes is

$$\widehat{x_{Aa}} = (x_{AA} + x_{Aa})\frac{2\hat{P_a}^{(0)}(1 - \hat{P_a}^{(0)})}{1 - (\hat{P_a}^{(0)})^2}.$$

We have just proportioned out the count of those with the dominant phenotype into the two underlying genotypes according to HW proportions.

Now, let's plug this back into Equation (3.3) as an estimate of x_{Aa}.

$$\hat{P_a}^{(1)} = \frac{1}{2n}\left[(x_{AA} + x_{Aa})\frac{2\hat{P_a}^{(0)}(1 - \hat{P_a}^{(0)})}{1 - (\hat{P_a}^{(0)})^2} + 2x_{aa}\right].$$

This equation can now serve as the basis of our iterative method. After we use this to calculate $\hat{P_a}^{(1)}$, we can plug that new number into the right side to get $\hat{P_a}^{(2)}$, and so on. We can more clearly show the iterative nature of this method by re-writing the above equation as

$$\hat{P_a}^{(j+1)} = \frac{1}{2n}\left[(x_{AA} + x_{Aa})\frac{2\hat{P_a}^{(j)}(1 - \hat{P_a}^{(j)})}{1 - (\hat{P_a}^{(j)})^2} + 2x_{aa}\right]$$

where $\hat{P_a}^{(j)}$ is the estimate of P_a at our j-th iteration, starting with the initial value where $j = 0$.

Now, as mentioned above, we will continue iterating until the change in our estimate from one iteration to the next is very small.

Example 3.4: Consider a two-allele locus where allele A is dominant to a. In a sample of 107 individuals, there are 85 of the dominant phenotype (this is $x_{AA}+x_{Aa}$) and 22 of the recessive phenotype (this is x_{aa}).

Our initial value calculation is

$$\hat{P}_a^{(0)} = \frac{2(22) + \frac{1}{2}(85)}{214} = .3014.$$

The table below shows the results of a number of iterations. Convergence happens on the 12th iteration at $\hat{P}_a = .45344$. Check these calculations for yourself. The best way is by setting up an Excel spreadsheet.

Iteration (j)	$\hat{P}_a^{(j)}$	$\widehat{x_{Aa}}$
0	0.30140	39.37
1	0.38959	47.66
2	0.42832	50.98
3	0.44383	52.26
4	0.44980	52.74
5	0.45207	52.93
6	0.45292	52.99
7	0.45325	53.02
8	0.45337	53.03
9	0.45341	53.03
10	0.45343	53.04
11	0.45344	53.04
12	0.45344	53.04

■

Example 3.5: **ABO Blood Group:** The ABO blood group provides a good example of using the E-M algorithm for a three-allele locus. In a homework problem, you will set the formulas up, but we'll set up the notation here.

First, we must recall the details of the alleles and genotypes at this locus. There are three alleles, labeled A, B, and O. Allele A is dominant to O, as is B. But A and B are codominant. The following tables summarize this and sets up a simpler notation to help work through this problem.

Allele	Population Frequency
A	P_A
B	P_B
O	P_O

Genotype	Phenotype	Count
AA		
	A	x_A
AO		
BB		
	B	x_B
BO		
AB	AB	x_{AB}
OO	O	x_O

■

Notice how important the HWE assumption is in this algorithm. The E-M algorithm could not be used here without it. If we are not sure about its validity, then we wouldn't be able to estimate such allele frequencies with any degree of accuracy.

3.5 Test for Hardy-Weinberg Equilibrium

In this section, we will formally discuss the procedure used to test for HWE. The background and setup of this test in the case of two alleles at a locus was discussed in the Overview section. The hypotheses are

$$\mathrm{H}_0 : \begin{cases} P_{AA} &= P_A^2 \\ P_{Aa} &= 2P_A(1-P_A) \\ P_{aa} &= (1-P_A)^2 \end{cases}$$

H_a :The above relationships do not hold

If we think carefully about this testing situation, we might realize that conducting this test is not really any different from the idea of the goodness-of-fit test of the preceding section. In other words, this test comes down to asking the question: "how well does our data fit the probability relationship described in the null hypothesis?" In fact, the section on conducting a goodness-of-fit test with a discrete distribution perfectly fits this situation.

More specifically, we can view the null hypothesis as

$$\mathrm{H}_0 : (X_{AA}, X_{Aa}, X_{aa}) \sim \text{Multinomial}\big[n, P_A^2, 2P_A(1-P_A), (1-P_A)^2\big]$$

We can view our data as coming in three categories, namely the three genotypic counts. We also notice that this null hypothesis is not completely specified since P_A is not known. So we'll have to estimate it using the gene counting MLE if possible, or otherwise the E-M algorithm. We can then compare observed and expected counts as before.

Let's look at the following MN blood group data from Hedrick (2000) based on a sample of 1000 individuals in England.

Genotype	Observed Prob.	Observed Count
MM	0.298	298
MN	0.489	489
NN	0.213	213

Our question is, is this locus in Hardy-Weinberg Equilibrium for the sampled population? The goodness-of-fit concept applied to this situation says that we should calculated the counts expected for each category as if the null hypothesis were true. To do that, we'll first need to estimate P_M (and we'll just take one minus this as our estimate of P_N). We can do that in this case using the simple gene counting estimate since the alleles are co-dominant and we have full genotypic information. The estimate is

$$\hat{P_M} = \frac{2 \times 298 + 489}{2 \times 1000} = 0.5425.$$

So our three expected probabilities are

$$(\hat{P_M})^2 = (.5425)^2 = 0.2943$$

$$2\hat{P_M}(1 - \hat{P_M}) = 0.4964$$

$$(1 - \hat{P_M})^2 = 0.2093$$

which allow us to expand our table to include expected counts as follows:

Genotype	Observed Prob.	Observed Count	Expected Prob.	Expected Count
MM	0.298	298	0.2943	294.3
MN	0.489	489	0.4964	496.4
NN	0.213	213	0.2093	209.3

Our test statistic follows the same concepts as previously with regard to its calculation and distribution. Since we have $k = 3$ categories in this data, the calculation here is

$$T = \sum_{j=1}^{3} \frac{(O_j - E_j)^2}{E_j} = \frac{(298 - 294.3)^2}{294.3} + \frac{(489 - 496.4)^2}{496.4} + \frac{(213 - 209.3)^2}{209.3} = 0.2215.$$

Under the null hypothesis, T will have a chi-squared distribution with $3 - 1 - 1 = 1$ degrees of freedom (since there are three categories, we always subtract one, and we estimated one parameter). So the p-value is (check on your own with Excel)

$$\text{p-value} = \text{P}(T > 0.2215) = 0.6379.$$

This is a large p-value, suggesting that the data appears to be very much in line with the null hypothesis. There is no evidence to reject the notion that this locus is in Hardy-Weinberg Equilibrium.

3.6 Fisher's Exact Test

We will now discuss a different and more complex testing methodology that is often used for tests of HWE. This generic procedure is referred to as an *exact test*, and the specific version of it that we will use for HWE test is called *Fisher's exact test* after the famous geneticist and statistician Sir Ronald Fisher.

The reason for considering a second method for conducting HWE tests goes back to our discussion regarding the validity of the chi-squared distribution. As we have mentioned, the test statistic we have been using has only approximately a chi-squared distribution, although that approximation is quite good if the sample is large (which is checked by looking at the expected counts). However, if the sample is not large, it is an inappropriate distribution to use for calculating the p-value.

We have also discussed a way to "fix things up" if the expected counts do not meet our requirements - namely, combining categories. However, this strategy is not a good one for HWE tests because it is not clear how to logically combine categories (i.e., genotypes), and the categories themselves are such an important part of the structure of Hardy-Weinberg proportions. Also, once we get into situations with more than two alleles, there can be a very large number of categories to deal with, which also makes it more likely that we will have a number of small expected counts.

Fisher's exact test is a method for calculating an accurate p-value for a test that doesn't rely on an approximate distribution. This is the meaning of the word "exact" in the name of the test; it calculates an exact p-value, not just an approximate p-value. It is often also referred to as a *probability test* for reasons we will see.

3.6.1 Fisher's Exact Test - Theory

A one sentence summary of the concept behind this procedure is: Given the allele counts that we observed, we will directly calculate the probability of our dataset having occurred, as well as any other datasets that *could have* occurred with the *same allele counts*.

Let's see how this works in the case of two alleles at the locus. We have seen that the appropriate probability models for genotype counts and allele counts if the null hypothesis is true (i.e., HWE is valid) are

$$(X_{AA}, X_{Aa}, X_{aa}) \sim \text{Multinomial}\big[n, P_A^2, 2P_A(1-P_A), (1-P_A)^2\big]$$

and

$$(X_A, X_a) \sim \text{Multinomial}(2n, P_A, 1-P_A)$$

respectively. And so the resulting joint pmfs are

$$\text{P}(X_{AA} = x_{AA}, X_{Aa} = x_{Aa}, X_{aa} = x_{aa}) = \frac{n!}{x_{AA}!x_{Aa}!x_{aa}!}(P_A^2)^{x_{AA}}\ [2P_A(1-P_A)]^{x_{Aa}}\ [(1-P_A)^2]^{x_{aa}}$$

and

$$\mathrm{P}(X_A = x_A, X_a = x_a) = \frac{(2n)!}{x_A! x_a!}(P_A)^{x_A}\ (1-P_A)^{x_a}$$

respectively.

Now, the conditional probability of observing genotype counts (X_{AA}, X_{Aa}, X_{aa}) given that the allele counts were (X_A, X_a) is

$$\begin{aligned}
\mathrm{P}(x_{AA}, x_{Aa}, x_{aa} \mid x_A, x_a) &= \frac{\mathrm{P}(x_{AA}, x_{Aa}, x_{aa})}{\mathrm{P}(x_A, x_a)} \\
&= \frac{\frac{n!}{x_{AA}! x_{Aa}! x_{aa}!}(P_A^2)^{x_{AA}}\ [2P_A(1-P_A)]^{x_{Aa}}\ [(1-P_A)^2]^{x_{aa}}}{\frac{(2n)!}{x_A! x_a!}(P_A)^{x_A}\ (1-P_A)^{x_a}} \qquad (3.4) \\
&= \frac{n! x_A! x_a! 2^{x_{Aa}}}{(2n)! x_{AA}! x_{Aa}! x_{aa}!}
\end{aligned}$$

Algebraically, you can notice that a lot of cancellations happen from the 2nd to 3rd line in Equation 3.4 by recalling that $x_A = 2x_{AA} + x_{Aa}$ and $x_a = 2x_{aa} + x_{Aa}$. Check this out for yourself.

So, Equation (3.4) is a conditional probability. It gives us the probability that we would observe genotype counts x_{AA}, x_{Aa}, and x_{aa} if it were known that the allele counts were x_A and x_a. Also note that the unknown allele frequencies have completely cancelled out of this formula. The probability calculation is only a function of observed counts. So, it is also, in fact, a potential test statistic. We will think of it and use it in both ways: as both a (conditional) probability calculation and as a test statistic.

3.6.2 Fisher's Exact Test - Methodology

Now we need to discuss how we make use of Equation (3.4) in our testing methodology. Below is a summary of the steps. Later, we will make it more clear through an example.

1. List out all the possible sets of genotype counts $(x^*_{AA}, x^*_{Aa}, x^*_{aa})$ that have the same allele counts (x_A, x_a) as observed in our dataset. Notice that one of these sets of genotype counts will be the same as the observed values (x_{AA}, x_{Aa}, x_{aa}).
2. For each of these possible genotype counts, apply Formula (3.4) to get the conditional probability that each would have occurred. If you have accounted for all possible sets of counts and made the calculations correctly, these probabilities will add to 1 (other than rounding error).
3. Sort this list from smallest probability to largest.
4. Find the row in this sorted table that pertains to the genotypic counts that were observed. The calculated p-value for the test, then, is the sum of that probability and all smaller probabilities.

If you follow the logic behind these steps, you will notice that the last step essentially makes the typical p-value calculation: the probability of being as extreme or more extreme than our observed data. "More extreme" is equated to our observed dataset and other less likely possible datasets.

Notice that this p-value will be small if our dataset was somewhat unusual (since it will have a small probability and there will be few datasets less likely), and will be larger if our dataset was somewhat typical (since it will have a larger probability and there will be many datasets less likely). So it follows our usual interpretation for a p-value.

3.6.3 Fisher's Exact Test - Example

Let's look at an example of applying Fisher's exact test. Suppose we have collected the following data on 20 individuals at one locus.

Genotype	Observed Count
AA	9
Aa	8
aa	3
	$n = 20$

For our observed data, the allele counts are $x_A = 2 \times 9 + 8 = 26$ and $x_a = 2 \times 3 + 8 = 14$ for a total of $2n = 40$ alleles.

Now we need to think of all the possible sets of genotype counts that would end up with these same allele counts. A systematic strategy for doing this would be to think of the case with the smallest possible number of heterozygotes as a starting point. Then another possible set of counts can be obtained by adding two heterozygotes to this and subtracting one homozygote. Continue this until we get to the point where it impossible to increase the heterozygote count and keep the allele counts as they need to be. This strategy is best understood by looking at this table, constructed from the example.

x^*_{AA}	x^*_{Aa}	x^*_{aa}
13	0	7
12	2	6
11	4	5
10	6	4
9	**8**	**3**
8	10	2
7	12	1
6	14	0

Notice that the boldface row represents the actual observed data.

The first row represents the situation with the smallest number of possible heterozygotes x_{Aa}, here 0. For this case, in order to have 26 A alleles and 14 a alleles, there must be 13 AA genotypes and 7 aa genotypes.

Continuing to the second row, we get this next possible set of counts by adding two to the x_{Aa} total above. Adding only one would leave us in a situation where there would be an odd total number of A and a alleles, and so we couldn't possibly have genotype counts with the required allele counts.

So, when we add two to the heterozygote count, we'll need to subtract one from each homozygote count in order to keep the allele counts the same. Keep doing this, creating new rows in the table, until we can't go any further. When we get down to the last row, notice that adding another row would require us to have a negative number for x_{aa} in order to keep allele counts the same, and of course this is not appropriate.

Now that we have a list of all possible genotype counts that have the observed allele counts, let's apply Equation (3.4) to each. Doing so gives us the following additional column.

x^*_{AA}	x^*_{Aa}	x^*_{aa}	Conditional Probability
13	0	7	.0000
12	2	6	.0006
11	4	5	.0145
10	6	4	.1070
9	**8**	**3**	**.3057**
8	10	2	.3669
7	12	1	.1779
6	14	0	.0274
			1.0000

Now let's sort this table from smallest to largest probability.

x^*_{AA}	x^*_{Aa}	x^*_{aa}	Conditional Probability
13	0	7	.0000
12	2	6	.0006
11	4	5	.0145
6	14	0	.0274
10	6	4	.1070
7	12	1	.1779
9	**8**	**3**	**.3057**
8	10	2	.3669
			1.0000

The resulting p-value is the sum of all probabilities less than or equal to the probability of our observed data. This is

$$\text{p-value} = .3057 + .1779 + .1070 + .0274 + .0145 + .0006 + .0000 = .6331.$$

The p-value for the test is .6331. This is large, and so there is no evidence that the null hypothesis is not correct. We believe that this locus seems to be in Hardy-Weinberg Equilibrium.

As a comparison, conduct this test on your own using the chi-squared goodness-of-fit methodology. You should wind up with a p-value of .5888. In this case, there is a difference in the two calculated p-values, but there is no difference in the final conclusion (luckily). If the p-values were of a smaller magnitude, such as in the .05 range, then such differences might be significant and lead to different conclusions. By the way, in conducting this as a chi-squared test, we would calculate the expected count for the *aa* genotype to be 2.45. Being one of just three categories and having an expected count less than 5, it leads us to believe that Fisher's exact test is the more appropriate methodology and leads to a more accurate p-value calculation.

As a second comparison, we could also use Fisher's exact test to compute a p-value for the MN blood group data we looked at earlier. The sample size was quite large in this case, so we would expect that the chi-squared approximate method would provide an accurate p-value. With such a large sample, it is impossible to create by hand a table like the one we just created, so we would need to rely on a computer. The resulting p-value based on Fisher's exact test is .6557, which is quite close to the p-value of .6379 computed using the chi-squared method. I used the DOS program at http://www2.biology.ualberta.ca/jbrzusto/hwenj.html to make this calculation. It can be used to apply Fisher's exact test to loci with any number of alleles, a situation which we'll introduce in the next section.

3.6.4 More than Two Alleles

Tests for HWE are commonly applied to situations where there are more than two alleles at the locus. The concept discussed above is the same, but the notation becomes a little more unwieldy since there are now many genotypes and alleles. We won't go through the formulas here, although you should be able to set up the models that underly them. As an example, if we have three alleles, A_1, A_2, and A_3, the null hypothesis would be

$$\begin{aligned} \mathrm{H}_0 :&(X_{A_1A_1}, X_{A_1A_2}, X_{A_1A_3}, X_{A_2A_2}, X_{A_2A_3}, X_{A_3A_3}) \\ &\sim \text{Multinomial}(n, (P_{A_1})^2, 2P_{A_1}P_{A_2}, 2P_{A_1}P_{A_3}, (P_{A_2})^2, 2P_{A_2}P_{A_3}, (P_{A_3})^2) \end{aligned}$$

The chi-squared version of the test for HWE would proceed exactly as before: compute estimates of the unknown allele frequencies, then compare observed and expected counts using the standard goodness-of-fit statistic. We would now be estimating two allele frequency parameters (remember the third is just one minus the sum of the first two), so the degrees of freedom for the chi-squared test statistic would be 6-1-2=3.

The more alleles there are, the more likely we will be in a situation where we will need to use Fisher's exact test instead of the chi-squared-based test. This is because there will be a larger number of genotypes, and so for a given sample size, the counts will be spread out thinner among them.

Computing the p-value using Fisher's exact test by hand becomes extremely difficult for three or more allele situations, so we can rely on the program like the one mentioned above.

As an example, use the goodness-of-fit test to practice the calculations for the following dataset on your own.

Genotype	Observed Count
A_1A_1	14
A_1A_2	22
A_1A_3	9
A_2A_2	10
A_2A_3	11
A_3A_3	3
	$n = 69$

Your should get p-value=.8484 using the chi-squared test, so it seems that this three-allele locus follows Hardy-Weinberg Equilibrium fairly closely. Was the chi-square test appropriate to use for this example?

Chapter 4

Linkage and other Two-Locus and Multi-Locus Analyses

4.1 Review of Linkage

The study of genetic *linkage* is the study of how close in physical proximity loci are to each other on a chromosome. The ultimate goal is to create physical or genetic maps of a chromosome (maps of where various marker loci and/or genes are on a chromosome and their physical relationship to each other), or to do disease gene mapping (narrow down the region of the genome in which a gene that contributes to a disease must be).

We will first be general in our discussion and consider any two loci, whether they represent a coding region (genes) or simply genetic markers (non-coding regions). Let the first locus, locus **A**, have two alleles in the population labeled A and a, and let the second be locus **B** with the two alleles B and b. We are interested in how far apart loci **A** and **B** are (if they are in fact on the same chromosome). In most situations, we can only determine this indirectly (and only approximately at that). In other words, we won't be able to pinpoint the precise locations of both of these loci in the genome, so we'll have to infer the distance between them.

One of the key concepts that we will use to help make this inference is crossovers that occur during meiosis, as we have discussed earlier in the course. Recall that the basic idea is that the closer two loci are to each other, the less likely a crossover is to happen between them during meiosis, and therefore a smaller percentage of resulting gametes will be recombinants. The further apart they are (or if they are actually on different chromosomes), the more likely one or multiple crossovers is likely to happen, leading to a higher percentage of gametes being recombinants (to a theoretical maximum of 1/2).

4.1.1 Measuring Distance

There are two common distance measurements that are used to describe the proximity of two loci **A** and **B**. One is called the *physical distance*. As implied by the name, this is the actual number of base pairs physically

between the two locations. It is typically measured in kilobases (kb) and notated by d_{AB}. Of the two, this is the more difficult to determine, and is rarely known precisely until a more complete physical map of that area of the chromosome is made.

The second is called the *genetic distance* between the loci. We'll notate this by x_{AB} (as in Weir, 1996). The units of this measurement is the centiMorgan (cM). The centiMorgan is a unit of map distance defined as the distance along which there is a 2% chance of at least one crossover occuring during a meiosis between **A** and **B**. So, the closer together two loci are, the fewer cM's between them. In the human genome, it turns out that 1cM is approximately equal to about 1 million bases (1000 kb). However, this is an average across the entire genome of 3.3×10^9 bases. It is known that this relationship can be very different in different regions. It has also been found that the relationship between 1cM and base pairs is very different across species, and even different for males and females in many species.

Most of our focus will be on this second measure, so let's say a little more about it. A related measure is the recombination fraction between loci **A** and **B**. This is defined as the percentage of gametes that will be recombinants as the result of many meioses in the population. The notation we will use is r_{AB}. Recall that this value must be in the range $[0, .5]$. The hard upper bound of 0.5 is due to the fact that even if there always is a cross-over between **A** and **B**, 1/2 of the gametes produced will still be of the non-recombinant type. More precisely, if an individual has genotype AB/ab, the gamete probabilities after many meioses are shown in Table 4.1 (as we saw earlier in the course).

Table 4.1: Gamete probabilities from recombination

		(recombinant)	(recombinant)	
Gamete	AB	Ab	aB	ab
Probability	$\frac{1-r_{AB}}{2}$	$\frac{r_{AB}}{2}$	$\frac{r_{AB}}{2}$	$\frac{1-r_{AB}}{2}$

Notice that the sum of the probabilities of the two recombinant gamete types is r_{AB}, as required. The above assumes that there are not double (or more) crossovers in the region, which is valid for short distances, but not for longer ones. Another result of this is that for short distances, 1% recombination is equal to 1cM.

4.2 Estimating r_{AB}

Our primary goal now will be to estimate the parameter r_{AB} for a pair of loci. A very direct way to estimate it precisely would be to observe the results of many meioses that occur for a particular individual. However, one should realize that this is a very invasive process (if we are talking about humans), and is also very time

consuming and difficult (for any species). We would have to monitor the microscopic process of meiosis for an individual, collect resulting gametes, determine the alleles on each gamete through a molecular process, then repeat this a large number of times for this and other selected individuals. This is just not a feasible process. In fact, the ability to determine the alleles on the gametes presupposes that we know where those loci are, which we don't - that is the whole reason for this problem in the first place.

So, the general experimental process that is used is to observe the offspring of matings between individuals. This is a reasonable substitute for directly observing a sample of meioses since the alleles (and gametes) passed on to offspring through mating are the result of those meioses we wanted to observe.

It should be noted, however, we still need to be aware of what we are capable (and incapable) of observing even in such an experiment. First, we still will not directly observe genotypes (for the most part) of these offspring since that would still require a lot of molecular work and presupposes that we know the locations of the loci. Instead, we will make use of an individual's phenotype, which may or may not directly infer their genotype.

Second, the above situation still poses a potential issue in many scenarios that we'll have to deal with. The problem is the following: If we don't know the precise genotype of the parental individuals (including the phase of the loci), phenotypic data will not help us infer whether or not the offspring was the result of a recombinant gamete. For example, if we notice an offspring with a doubly recessive phenotype, we know the genotype must be ab/ab. However, the ab gametes could be a recombinant type (if the parent was Ab/aB for example) or a non-recombinant type (if the parent was AB/ab for example). So without knowing the phase of doubly dominant parents, we won't be able to infer whether an offspring's gametes were recombinant or not through their phenotype. Fortunately, we can set things up experimentally to help overcome this problem, as we'll see.

4.2.1 Backcross Experiment

If our study is on plants or certain animals, we can setup mating experiments in a way that we will be able to infer recombinants precisely. First, let's make it clear that the each loci we are studying must have a dominant-recessive relationship (with the capital letter representing the dominant allele).

The first experiment type we can perform is called a *backcross experiment*. In this experiment, we would start off with genetic lines that are completely homozygous at both loci - some for the dominant alleles (AB/AB) and some for the recessive alleles (ab/ab). Crossing these two lines (i.e., allowing matings to occur between individuals from one line to individuals from the other line) produces a population of individuals called the F_1 population. All individuals in F_1 will be doubly heterozygous (AB/ab) with phase known.

Now this F_1 population of individuals is *backcrossed* with the doubly recessive line (ab/ab) of individuals. The offspring in the resulting F_2 population will always receive an ab gamete from the doubly recessive

parent, but the doubly heterozygous parent will provide one of four possible gametes to offspring: AB, Ab, aB, or ab. The second and third are recombinant types.

So, the offspring of the backcross will fall into one of four possible genotypic categories. And, due to the setup of the experiment, each of these four genotypes will outwardly display a different phenotype, and so the underlying genotype can be directly inferred. In addition, the percentage of offspring we would expect to fall into each of these categories can be written based on the recombination fraction r_{AB}. Table 4.2 summarizes this.

Table 4.2: Backcross genotypic probabilities

Genotype	Phenotype	Probability
AB/ab	Dominant at **A**, Dominant at **B**	$(1 - r_{AB})/2$
Ab/ab	Dominant at **A**, Recessive at **B**	$r_{AB}/2$
aB/ab	Recessive at **A**, Dominant at **B**	$r_{AB}/2$
ab/ab	Recessive at **A**, Recessive at **B**	$(1 - r_{AB})/2$

Our goal now is to estimate the unknown r_{AB} based on resulting data from such an experiment. Luckily, this is not difficult because we notice that this setup falls nicely into a multinomial distribution situation. We have a random sample of individuals that each fall into one of four categories, whose probabilities can be written in terms of our unknown parameter. In fact, we can reduce this to a binomial distribution because all we are really interested in is whether or not a recombination event occurred between **A** and **B**. So, the first and fourth categories can be combined and called the non-recombinant (or parental) category, and the second and third categories can be combined and called the recombinant category in our experiment, as summarized in Table 4.3.

Table 4.3: Backcross as a binomial experiment

Binomial Category	Probability	Observed Count
Recombinant	r_{AB}	n_R
Parental	$1 - r_{AB}$	n_P

We will observe n total individuals and count the number n_R of recombinant types, and the remaining will be n_P. We can now model the random variable n_R as a binomial random variable.

$$n_R \sim \text{Bin}(n, r_{AB})$$

This is exactly the same setup as our standard binomial distribution, with r_{AB} playing the role of p in the generic notation. So, all of our estimation procedures work just as before. In other words, the MLE for r_{AB} is the sample proportion that fell into the recombinant category:

$$\widehat{r_{AB}} = n_R/n.$$

There is one slight adjustment we will make to this estimator to account for the fact that we know the largest possible value for r_{AB} is 0.5 based on the known constraint of meiosis. It is possible due to sampling variation that we will observe the proportion of recombinants to be greater than 0.5 in our experiment. If this is the case, we will use 0.5 as our estimate of r_{AB} instead of n_R/n. So, we can more precisely write the estimator as

$$\widehat{r_{AB}} = \min(n_R/n, 0.5).$$

As we have seen before, the estimator $\widehat{r_{AB}}$ is an unbiased estimator for r_{AB} and has standard deviation

$$\sigma_{\widehat{r_{AB}}} = \sqrt{\frac{r_{AB}(1-r_{AB})}{n}}$$

An approximate 95% confidence interval for r_{AB} can then be calculated as

$$\widehat{r_{AB}} \pm 2 \times \sqrt{\frac{r_{AB}(1-r_{AB})}{n}}.$$

Example 4.1: The following data was reported for maize by Goodman (1980). The two loci **A** and **B** represent different enzyme-producing genes in the plant. Following the experimental backcross method, the following genotype (and phenotype) counts were observed:

Genotype	Observed Count
AB/ab	53
Ab/ab	13
aB/ab	9
ab/ab	55

The total sample size was $n = 130$ and $n_R = 22$ was the observed number of recombinants. The maximum likelihood estimate of the recombination frequency between these loci is

$$\widehat{r_{AB}} = 22/130 = 0.1692$$

with estimated standard deviation

$$\sigma_{\widehat{r_{AB}}} = \sqrt{\frac{0.1692(1-0.1692)}{130}} = 0.0329$$

Our approximate 95% confidence interval for r_{AB} is

$$0.1692 \pm 2 \times 0.0329 = (0.1034, 0.2021).$$

This is a fairly wide range, due to the sample size for the experiment.

■

We may also want to conduct a hypothesis test to ask the question of whether the two loci are linked or not (in addition to, or instead of, estimating the value of the recombination fraction). This can be accomplished simply with a chi-square goodness-of-fit test with the following hypotheses

$$\text{H}_0 : \textbf{A} \text{ and } \textbf{B} \text{ are not linked (i.e., } r_{AB} = 0.5 \text{)}$$
$$\text{H}_a : \textbf{A} \text{ and } \textbf{B} \text{ are linked (i.e., } r_{AB} < 0.5 \text{)}$$

The frequencies we would expect in the two categories under the null hypothesis are 1/2 and 1/2, so we can construct our table as in Table 4.4.

Table 4.4: Observed and expected counts for linkage test

Category	Observed Count	Expected Count
Recombinant	n_R	$n/2$
Parental	$n_P = n - n_R$	$n/2$

The test will have one degree of freedom since there are two categories and the null hypothesis has no unknown parameters.

Example 4.2: To continue the maize example, we can conduct the chi-square test by constructing the following table of observed and expected counts

Category	Observed Count	Expected Count
Recombinant	22	130/2=65.5
Parental	108	130/2=65.5

The test statistic is $T = \frac{(22-65.5)^2}{65.5} + \frac{(108-65.5)^2}{65.5} = 57.78$. This has a very small p-value (zero to more than four decimal places) for a χ^2_1 distribution, and so we clearly reject the null hypothesis. The loci seem to be linked.

■

4.2.2 F_2 Experiment

Another experiment that can be used in plants is called an F_2 *experiment*. In this experiment, an F_1 population is created just as with the backcross experiment. This population has all doubly heterozygous individuals with phase known. Now, however, this population is allowed to mate amongst themselves to create an F_2 population. When double heterozygotes mate, there will be nine possible resulting genotypes

with four noticeable phenotypes. We want to summarize these along with their expected probabilities.

To do this, first notice that each individual in such a mating produces four possible gametes in proportions shown in Table 4.5.

Table 4.5: Gamete proportions for F_2 experiment

Gamete	Probability
AB	$(1-r_{AB})/2$
Ab	$r_{AB}/2$
aB	$r_{AB}/2$
ab	$(1-r_{AB})/2$

Now, the potential resulting offspring genotypes can be viewed by considering which gamete was received from each parent. This is displayed in Table 4.6. Notice that the phase of the offspring will not necessarily be known, but we can write down the possible phase-unknown genotypes. This is enough to allow us to connect it to the observable phenotypes. Table 4.7 shows the probabilities corresponding to each genotype in Table 4.6.

Table 4.6: Resulting offspring genotypes from F_2 experiment

	Parent 1			
Parent 2	AB	Ab	aB	ab
AB	$AABB$	$AABb$	$AaBB$	$AaBb$
Ab	$AABb$	$AAbb$	$AaBb$	$Aabb$
aB	$AaBB$	$AaBb$	$aaBB$	$aaBb$
ab	$AaBb$	$Aabb$	$aaBb$	$aabb$

Table 4.7: Resulting offspring probabilities, corresponding to genotypes in Table 4.6

	Parent 1			
Parent 2	AB	Ab	aB	ab
AB	$(1-r)^2/4$	$r(1-r)/4$	$r(1-r)/4$	$(1-r)^2/4$
Ab	$r(1-r)/4$	$r^2/4$	$r^2/4$	$r(1-r)/4$
aB	$r(1-r)/4$	$r^2/4$	$r^2/4$	$r(1-r)/4$
ab	$(1-r)^2/4$	$r(1-r)/4$	$r(1-r)/4$	$(1-r)^2/4$

Because of the symmetry of this genotype matrix, there are 9 (not 16) different genotypes. They will correspond to 4 observable phenotypes as shown in Table 4.8.

Table 4.8: Correspondence of genotypes to phenotypes in F_2 experiment

Genotype	Phenotype	Phenotype Notation
$AABB$	Dominant A, Dominant B	AB
$AaBB$	Dominant A, Dominant B	
$AABb$	Dominant A, Dominant B	
$AaBb$	Dominant A, Dominant B	
$AAbb$	Dominant A, Recessive B	Ab
$Aabb$	Dominant A, Recessive B	
$aaBB$	Recessive A, Dominant B	aB
$aaBb$	Recessive A, Dominant B	
$aabb$	Recessive A, Recessive B	ab

With this mapping of phenotypes to genotypes in mind, we can now write down the probability model for the observed phenotype counts (recall, all we will be able to observe is phenotypes, not genotypes). Clearly, the multinomial model with $k = 4$ is appropriate. We just need to figure out the expected probability for each category by adding up the probabilities from the appropriate cells of the table above. We will do this below for just one of the four categories, but you should go through the process on your own for each of them.

Let's focus on the phenotype Ab. Tables 4.9 and 4.10 reproduce Tables 4.6 and 4.7, but with the related cells highlighted.

Table 4.9: Resulting offspring genotypes from F_2 experiment (highlight Ab phenotype)

	Parent 1			
Parent 2	AB	Ab	aB	ab
AB	$AABB$	$AABb$	$AaBB$	$AaBb$
Ab	$AABb$	$\boxed{AAbb}$	$AaBb$	$\boxed{Aabb}$
aB	$AaBB$	$AaBb$	$aaBB$	$aaBb$
ab	$AaBb$	$\boxed{Aabb}$	$aaBb$	$aabb$

We can see that the probability of this phenotype is the sum of the highlighted probabilities in the table. This is

$$\frac{r^2}{4} + \frac{2r(1-r)}{4} = \frac{2r - r^2}{4} = \frac{r(2-r)}{4} = \frac{1-(1-r)^2}{4}$$

The algebraic manipulation leading to the last equality above has been made for a reason we will see below. Going through this process for the other three categories gives us the probabilities in Table 4.11.

The reason for writing each of these probabilities in terms of the factor $(1-r)^2$ is so that we can make a substitution that will make the problem easier. Letting $\theta = (1-r)^2$ and substituting into the probabilities

Table 4.10: Resulting offspring probabilities, corresponding to genotypes in Table 4.9 (highlight Ab phenotype)

	Parent 1			
Parent 2	AB	Ab	aB	ab
AB	$(1-r)^2/4$	$r(1-r)/4$	$r(1-r)/4$	$(1-r)^2/4$
Ab	$r(1-r)/4$	$\boxed{r^2/4}$	$r^2/4$	$\boxed{r(1-r)/4}$
aB	$r(1-r)/4$	$r^2/4$	$r^2/4$	$r(1-r)/4$
ab	$(1-r)^2/4$	$\boxed{r(1-r)/4}$	$r(1-r)/4$	$(1-r)^2/4$

Table 4.11: Phenotype probabilities from F_2 experiment

Phenotype	Probability
AB	$\frac{2+(1-r)^2}{4}$
Ab	$\frac{1-(1-r)^2}{4}$
aB	$\frac{1-(1-r)^2}{4}$
ab	$\frac{(1-r)^2}{4}$

from Table 4.11 gives us the simplified probability forms in Table 4.12.

Table 4.12: Phenotype probabilities from F_2 experiment, written in terms of $\theta = (1-r)^2$

Phenotype	Probability
AB	$\frac{2+\theta}{4}$
Ab	$\frac{1-\theta}{4}$
aB	$\frac{1-\theta}{4}$
ab	$\frac{\theta}{4}$

These probabilities are a little simpler looking and will be easier to deal with. Also, even though our ultimate goal is to find an estimator for r, we can do that by first finding an estimator for θ (call it $\hat{\theta}$), and then our estimator for r will be $\hat{r} = 1 - \sqrt{\hat{\theta}}$ by just writing r in terms of θ.

Now, with all this behind us, we can fully state our probability model. We have four observed counts, call them X_{AB}, X_{Ab}, X_{aB}, and X_{ab} with the following multinomial model.

$$(X_{AB}, X_{Ab}, X_{aB}, X_{ab}) \sim \text{Multinomial}\Big(n, \frac{2+\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\Big)$$

The joint pdf of these random variables, and therefore the likelihood function for θ is

$$\text{L}(\theta) = \frac{n!}{X_{AB}!X_{Ab}!X_{aB}!X_{ab}!}\Big(\frac{2+\theta}{4}\Big)^{X_{AB}}\Big(\frac{1-\theta}{4}\Big)^{X_{Ab}}\Big(\frac{1-\theta}{4}\Big)^{X_{aB}}\Big(\frac{\theta}{4}\Big)^{X_{ab}}$$

Notice that we just have one parameter in our model (θ). Also notice that this multinomial model is much

more complex than any we have come across before (if you don't notice it now, it will become very apparent shortly).

To find the MLE for θ, we need to go through the usual steps. First, the log-likelihood is

$$\ln \mathrm{L}(\theta) = C + X_{AB} \ln\Big(\frac{2+\theta}{4}\Big) + X_{Ab} \ln\Big(\frac{1-\theta}{4}\Big) + X_{aB} \ln\Big(\frac{1-\theta}{4}\Big) + X_{ab} \ln\Big(\frac{\theta}{4}\Big)$$

and the derivative with respect to θ is

$$\frac{\partial \mathrm{L}(\theta)}{\partial \theta} = \frac{X_{AB}}{2+\theta} - \frac{X_{Ab}+X_{aB}}{1-\theta} + \frac{X_{ab}}{\theta}$$

Now we need to set this equal to zero and solve for θ. This process is not nearly as easy as it has been in other examples to this point, but it can be done. First, we'll multiply through by the least common denominator and work towards solving for θ.

$$\begin{aligned}
0 &= \theta(1-\theta)X_{AB} - \theta(2+\theta)X_{Ab} + X_{aB}) + (1-\theta)(2+\theta)X_{ab} \\
0 &= \theta X_{AB} - \theta^2 X_{AB} - \theta^2(X_{Ab}+X_{aB}) - 2\theta(X_{Ab}+X_{aB}) + 2X_{ab} + \theta X_{ab} - 2\theta X_{ab} \\
0 &= -\theta^2(X_{AB}+X_{Ab}+X_{aB}+X_{ab}) + \theta(X_{AB} - 2[X_{Ab}+X_{aB}] - X_{ab}) + 2X_{ab} \\
0 &= n\theta^2 - (X_{AB} - 2X_{Ab} - 2X_{aB} - X_{ab})\theta - 2X_{ab}
\end{aligned}$$

This is now a quadratic equation in θ which can be solved using the quadratic formula. We get

$$\theta = \frac{(X_{AB} - 2X_{Ab} - 2X_{aB} - X_{ab}) \pm \sqrt{(X_{AB} - 2X_{Ab} - 2X_{aB} - X_{ab})^2 + 8nX_{ab}}}{2n}$$

It is easy to see that using the negative sign version of the above solution leads to a negative number, which is an inappropriate value for θ. Therefore, our MLE for θ is the positive sign version of the above, giving us

$$\hat{\theta} = \frac{(X_{AB} - 2X_{Ab} - 2X_{aB} - X_{ab}) + \sqrt{(X_{AB} - 2X_{Ab} - 2X_{aB} - X_{ab})^2 + 8nX_{ab}}}{2n}$$

Recall that our real goal was to estimate r which is a function of θ. This estimator is

$$\hat{r} = 1 - \sqrt{\hat{\theta}}$$

The calculation of the variance of the estimator $\hat{\theta}$ is not easy and the technique is beyond the scope of our discussion. However, the formula is below

$$\sigma^2_{\hat{\theta}} = \frac{2\theta(2+\theta)(1-\theta)}{n(1+2\theta)}$$

where we would replace θ by $\hat{\theta}$ in the formula when calculating it.

Example 4.3: Suppose we data from an F_2 experiment where 1929 plants were categorized into the four phenotypic categories as follows

Phenotype	Observed Count
AB	1221
Ab	219
aB	246
ab	243

We wish to estimate the the recombination fraction between the loci. Our estimate of θ using the above formula is

$$\hat{\theta} = \frac{48 + \sqrt{48^2 + 3749976}}{3858} = 0.5145.$$

This leads to an estimate for r of

$$\hat{r} = 1 - \sqrt{\hat{\theta}} = 1 - \sqrt{0.5145} = 0.2827.$$

We would say, then, that we estimate that 28% of the gametes produced by meiosis in this species are of the recombinant type with regard to these two loci. For now, we can also say that the distance between the loci is 28cM.

■

4.2.3 Human Pedigree Analysis

For analysis of linkage in humans, clearly we cannot conduct experiments such as those discussed above. The general method used is to analyze family *pedigrees*, which is simply a "family tree" type of picture that shows relationships among individuals in a family as well as known genetic information about each. Very often, such analyses are conducted with the goal of finding the location (or at least narrowing down to a small area) of genes that cause diseases.

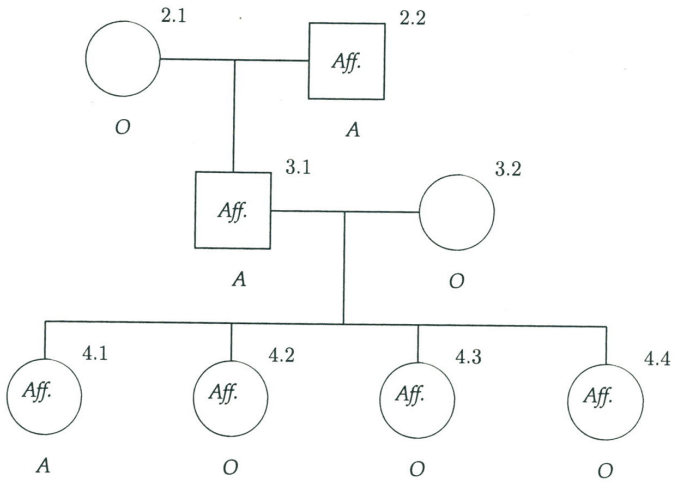
The basic idea is fairly simple. It is really a matter of getting familiar with the notation in a pedigree and careful analysis of the passing of genes from one generation to the next.

The sample pedigree in Figure 4.1 (a fairly simple one) is from Weir (1996) and represents a family with individuals who have hereditary motor and sensory neuropathy type 1 (see

http://www.emedicine.com/neuro/byname/
charcot-marie-tooth-and-other-hereditary-motor-and-sensory-neuropathies.htm

for more info on this disease of the muscular system, usually in children). This disease gene is referred to as *CMT1*, and is dominant, meaning that affected individuals have at least one dominant allele. Normal individuals are homozygous for the recessive allele. We'll use Weir's notation and refer to the alleles for this

Figure 4.1: Sample family pedigree from Weir (1996)



gene as T and t. Affected individuals have genotype Tt and normal individuals are tt.

In pedigrees, squares represent males and circles represent females. Typically, individuals who are affected with the disease in question have their symbol darkened in, but in this example, they have the notation *Aff.* within their symbol.

The other locus for which information is represented here is the ABO blood group. The notation A or O under each individual represents their blood type (a phenotype). Apparently, no individuals in this pedigree have blood type B or AB.

So, our goal is to estimate the recombination fraction r between the *CMT1* disease locus and the ABO blood group locus. Since the location of ABO is known, if we find that *CMT1* is close to it (i.e., small r), it can help narrow down our search for the disease gene. We will make an estimate from this pedigree by trying to detect recombination events that must have occurred as genes were passed down to the four children in the most recent generation.

In our analysis of the pedigree, we must carefully walk through individual by individual and attempt to infer what their phase known genotypes must be based just on their known phenotypes. We will first focus on the individual labeled 3.1. This person is Tt for the disease gene and is at first glance either AA or AO for the ABO gene. However, since his spouse is blood type O, she must be OO at that gene. Therefore, 3.1 cannot be AA because some of the children of 3.1 are O blood types, so they must have received an O allele from each parent. We now know that 3.1 must be AO. So the phase-unknown genotype is $Tt\ AO$.

We can deduce the correct phase, however, by looking at his parents. Individual 2.1, being unaffected and being of blood type O, must have phase-known genotype tO/tO. This means that 3.1 will have received a

gamete containing the alleles tO from this parent. We can now say that 3.1's phase known genotype is TA/tO.

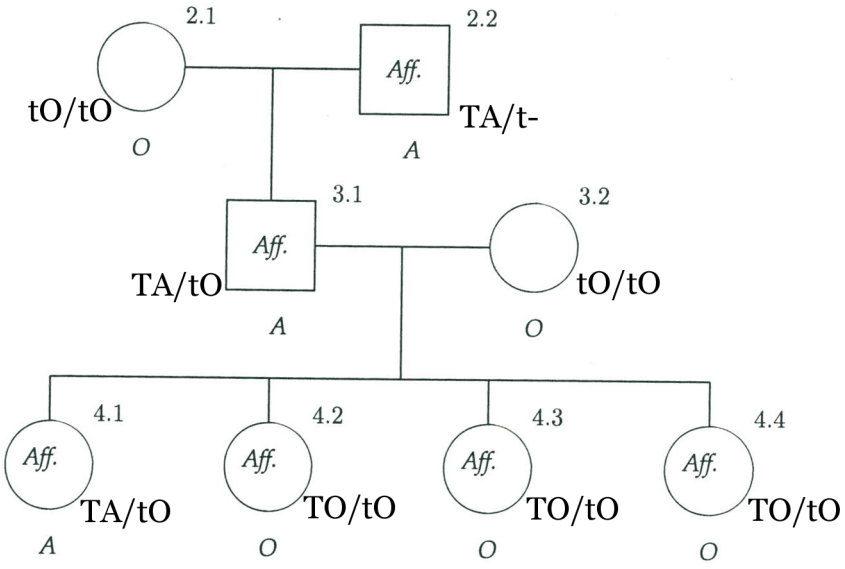
For individual 3.2, like individual 2.1, the genotype must be tO/tO. So children of this 3.1-3.2 mating will always get a tO gamete from their mother. We will not be able to make any recombination estimates based on 3.2's passing on of genetic information. But we can do so for 3.1.

Now, looking at the children of this mating, they are all affected and so always got the T allele from their father (since their mother doesn't have that allele to pass on). Notice that the children 4.2, 4.3, and 4.4 are all O blood types. So they must also have received an O allele from their father. Therefore, 3.1 passed on the gamete TO to these three children, which is a recombinant gamete since his genotype is TA/tO.

Finally, child 4.1 has A blood type, so must have received the non-recombinant gamete TA from 3.1.

So, we count three recombination events out of four that we can analyze. 75% of the observable meioses in this pedigree resulted in recombinant gametes. Our estimate of r then is 0.5 (it would be .75, but we have seen that .5 is our hard upper bound for values of r). This was a very small pedigree and a very small amount of data to come up with accurate estimates. But at first glance, it may seem that these two loci are not tightly linked.

Below is the pedigree again, with our analysis of genotypes included.



4.2.4 Genetic Markers

One tool that has helped geneticists locate positions of genes along a chromosome is genetic markers. In general, we can think of a marker as a special locus (i.e., a location along a chromosome) which is not in a coding region; in other words, the associated DNA sequence is not transcribed into a protein. There is a

lot to still be learned about these non-coding regions of DNA (which make up a very high percentage of an individual's DNA), but the presence of markers in these areas is very helpful.

More specifically, markers are regions of DNA that can be detected based on the particular sequence of bases that it contains. One very common example are *restriction enzyme fragments*. These are sequences of DNA that are detected by a certain enzyme. For example, and enzyme named *HindIII* is found in bacteria and when allowed to react with a DNA molecule, it will "find" the sequence of bases AAGCTT (in the 5' to 3' direction) and clip the DNA after the first A, creating fragments in the original sequence. Other enzymes recognize other base sequences. For example, the enzyme *EcoRI* recognizes the sequence GAATTC and clips the DNA after the G. The enzyme *HpaI* recognizes GTTAAC and clips the DNA after the second T. There are hundreds of known such enzymes, each recognizing different DNA sequences. The resulting DNA fragments are called *restriction fragments*, and their size and number within a particular DNA sequence can be used to differentiate between individuals (i.e., can be considered different alleles for this marker locus).

As an example, consider an individual who has the following two complimentary strands of DNA along some chromosome:

```
AATTCCAGGTTAACCCCAGGTTTTCAGGTTGTTAACCAAGGTTTGGA
AATTCCGAGTTAACCCCAGTTTTGCAGGGGTTACTAGCCCAGTTCAT
```

If we allow the enzyme *HpaI* to "digest" the first strand above, notice that it will find two GTTAAC sequences (its recognition sequence) and clip the DNA after the second T. The result is three DNA fragments:

```
AATTCCAGGTT    AACCCCAGGTTTTCAGGTTGTT    AACCAAGGTTTGGA
```

When it digests the second strand, it finds only one GTTAAC sequence, and so the result is only two fragments:

```
AATTCCGAGTT    AACCCCAGTTTTGCAGGGGTTACTAGCCCAGTTCAT
```

So, for this particular location in the genome, we might consider these two situations as two different possible alleles, described by the size and number of fragments left after being digested by *HpaI*. We might consider the first digestion result allele 1 and the second allele 2. A different individual might have a DNA sequence at this location that leads to a different pattern of fragments which we might label allele 3, and so on.

It should be noted that the laboratory process to actually do everything stated above is not a simple process. A lot of work goes into recognizing the number and size of fragments created (remember, these are very small, microscopic molecules, so it is not as simple as just looking down and seeing the three fragments).

These restriction fragment sites are just one example of the kind of marker site that can be detected through laboratory analysis. Other examples are sites that have repeats of the same sequence many times, called

VNTR (Variable Number of Tandem Repeats) or STR (Short Tandem Repeats) sites, RAPD Loci (Random Amplified Polymorphic DNA), and SNPs (Single Nucleotide Polymorphisms).

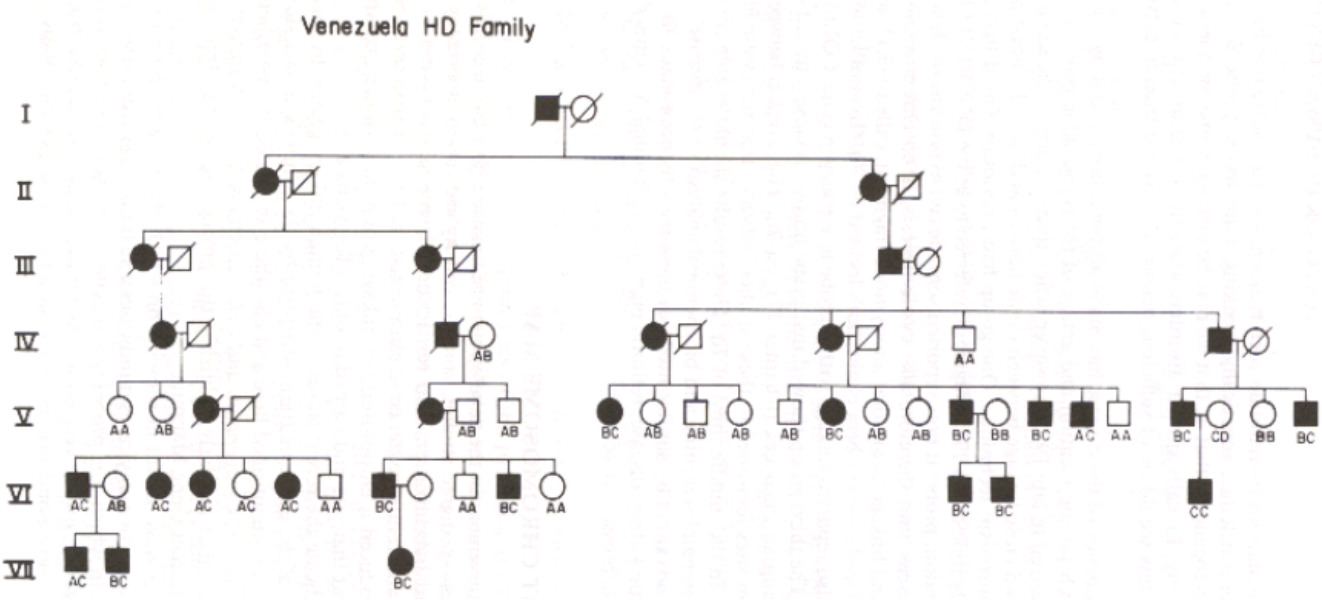
4.2.5 Genetic Markers in Disease Analysis

In the human pedigree example we saw above, we attempted to determine the distance between a disease gene and the ABO blood group locus. We can use loci such as ABO or a few other well known and understood genes to complement these pedigree analyses. The downside is that such loci are few and far between. So, if we really want to narrow our search area for a disease gene down to a relatively small region of DNA, it won't help us much.

That is where these genetic markers come in. Markers such as restriction fragment sites, VNTR's etc., are found in great quantities all over the human genome, and so they can be very useful in narrowing down the search for the location of a gene. As a general method, we can keep measuring recombination fractions between the disease gene and many such marker loci until we begin finding markers that tend to have low recombination rates. Then we know we are nearing the true location of the gene. This is the basic method that was used to find genes such as for Cystic Fibrosis. A lot more is known about this disease now that its location was found and its DNA sequence and related protein sequence has been analyzed.

As an example, Figure 4.2 shows a pedigree of a family where Huntington's disease is prevalent. Huntington's is a dominant disease, so that, as with the previous example, the homozygous recessive genotype *hh* is the normal condition.

Figure 4.2: Huntington's disease pedigree



This pedigree is quite complete, going back seven generations. For individuals in the recent generations who were still alive (a slash through an individual's symbol means they have died), the notations AA,AB,BC, etc under their symbol refer to the genotype of a restriction fragment site. This particular restriction site has four different alleles, which are labeled A, B, C, and D.

This pedigree is just included here as an example of the use of genetic markers to construct linkage analyses involving disease genes. Although this particular pedigree is large, its usefulness for detecting recombination events is not great because many of the parents of the large families in the recent generations died before being able to be genotyped for the restriction fragment site.

4.3 Constructing Genetic Maps

We will now make use of estimated recombination fractions between loci to construct genetic maps of chromosomes.

4.3.1 Relationship between Map Distance and Recombination Frequency

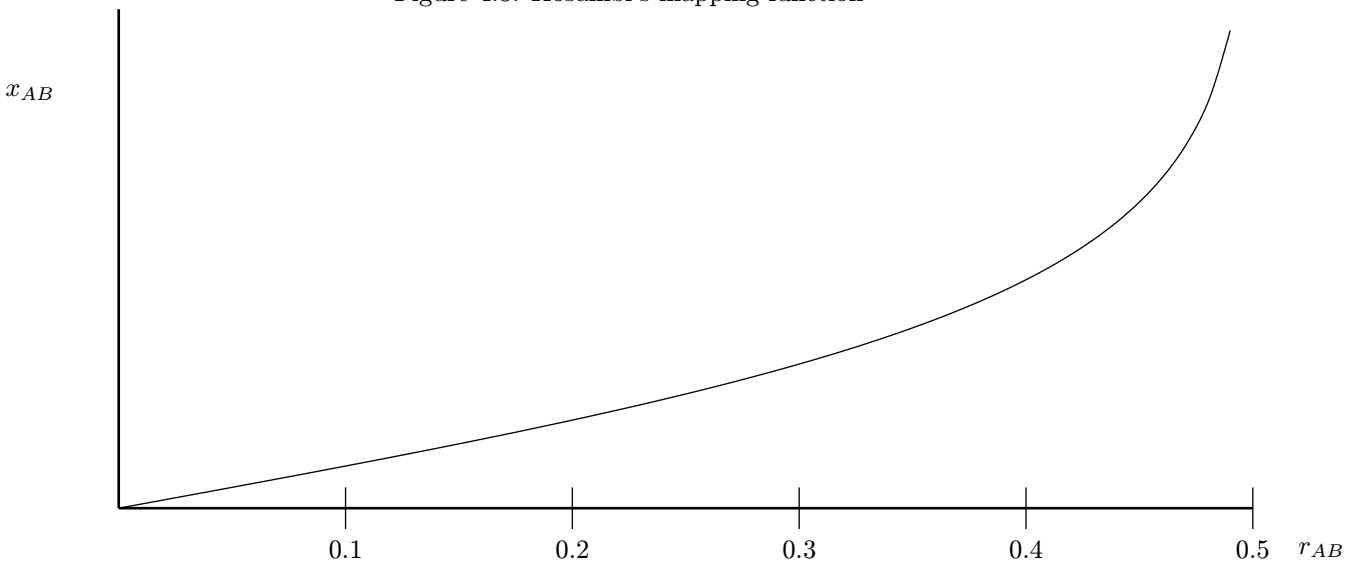
We noted earlier that genetic distance (also called *map distance*) is the typical measure used to identify distances between loci along a chromosome, and so is used often in constructing genetic maps. Recall that genetic distance is measured in centiMorgans (cM), where one cM is defined as the length of chromosome along which we expect a 2% chance of at least one crossing-over occurring during a meiosis. A centiMorgan is also referred to as a map unit (1cM = 1 map unit).

It may at first glance seem that the centiMorgan and recombination frequency are the same thing since we have stated earlier that 2% crossing-over equal 1% recombination. However, notice that the definition of a centiMorgan involves the probability of "crossing-over occurring", while the definition of recombination frequency involves the percentage of recombinant gametes produced. These two are not the same if an *even* number of crossovers occurred between the loci during the meiosis (i.e., $r < c/2$ when there can be two or more crossovers). If that were the case, crossing-over would have occurred, but recombinants would not have been produced. For very short distances, we don't have to worry about this since two (or more) crossovers within a small region is rare (this fact is called *interference*). But for greater distances, it is more common for two or more crossovers to occur, and so the map unit and recombination frequency are no longer equivalent.

Since we construct genetic maps in map units, but we make measurements of recombination frequency from experimental data and pedigrees, we first need to be sure that we are getting accurate measurements of map units based on recombination data. One way to do this is with a *mapping function*. There are many such mapping functions that have been suggested. One that we'll discuss is a function derived by Kosambi in 1944 (we'll call it the Kosambi mapping function). This function relates recombination frequency to map distance in the following way (recall that x_{AB} is our notation for the map distance, in cM, between loci **A** and **B**), and plotted in Figure 4.3.

$$x_{AB} = \frac{1}{4}\ln\Big(\frac{1+2r_{AB}}{1-2r_{AB}}\Big) \tag{4.1}$$

Figure 4.3: Kosambi's mapping function



Notice that this function is linear in the regions of small values of r_{AB}. This takes into account that double crossovers are rare over short distances, meaning map distance and recombination frequency are nearly the same. But for larger values of r_{AB}, the two measures begin diverging, and map distance is greater than recombination frequency. In this function x_{AB} ranges from 0 to ∞ as r_{AB} ranges from 0 to 0.5.

As an example, in our F_2 plant experiment earlier in this section, we estimated $r_{AB} = .283$. At the time, we said that the map distance could be estimated as 28.3cM. However, using Kosambi's formula, we can get a better estimate of map distance. The calculation for that example would be

$$x_{AB} = \frac{1}{4}\ln\Big(\frac{1+2(.283)}{1-2(.283)}\Big) = .321$$

So we would estimate the map distance between **A** and **B** more accurately as 32.1cM.

It should be noted again that this mapping function is just one of many that have been suggested. It should not be considered perfect, and works better in some situations than others.

Another way to use recombination frequency to accurately determine map distance between two loci **A** and **B** is through a more exacting experimental method called a three-point testcross (or called three-point backcross). The idea is to expand our experiment so that we can directly observe double crossovers between **A** and **B**. To see how it will work, let's first suppose we did a standard backcross as we discussed earlier

(we might more accurately call that a two-point backcross now). Let's say that out of 100 individuals, the offspring had the genotypes (inferred from phenotypes) given in Table 4.13.

Table 4.13: Example genotypic counts for a backcross

Genotype	Count
AB/ab	37
Ab/ab	10
aB/ab	13
ab/ab	40

We would use this data and our simple estimation procedure for backcrosses to estimate $r_{AB} = (10 + 13)/100 = .23$. We might be tempted to claim that the distance in map units is 23cM. However, it is certainly possible that double crossovers occurred in some of these meioses that would have resulted in parental types being transmitted even though crossover events were happening.

To help detect such double crossovers, we can use a third locus (probably a marker locus) that we know to be between the other two. Let's call this third locus **C**. Now a three-point backcross experiment can be run. In this experiment, we will cross triply heterozygous *ACB/acb* individuals with homozygous recessive *acb/acb* individuals and observe the results. In this case, we will be able to distinguish eight phenotypic and associated genotypic classes as shown in Table 4.14.

Table 4.14: Three-point backcross relationship between genotypes and parental/recombinant types

ACB/acb *acb/acb*	Parental Types (No crossover occurred)
Acb/acb *aCB/acb*	Recombinant Types (One crossover between **A** and **C**)
ACb/acb *acB/acb*	Recombinant Types (One crossover between **C** and **B**)
AcB/acb *aCb/acb*	Recombinant Types (Two crossovers: between **A**,**C** and **C**,**B**)

The single crossover types were already detected by the two-point backcross. It is the double crossovers that are newly detected. Some of the original gametes that we labeled as parental types in the two-point backcross can now be seen to be recombinant types where a double crossover occurred. Our new data might be as in Table 4.15.

We can now use this to estimate the recombination frequencies r_{AC} and r_{CB}. These estimates from this data are

$$r_{AC} = \frac{5+8+3+4}{100} = .20 \quad \text{and} \quad r_{CB} = \frac{5+5+3+4}{100} = .17$$

So the map distance between **A** and **C** is 20cM and between **C** and **B** is 17cM. Map distances are additive

Table 4.15: Example genotypic counts for three-point backcross (extension of Table 4.13)

Genotype	Count
ACB/acb	34
acb/acb	36
Acb/acb	5
aCB/acb	8
ACb/acb	5
acB/acb	5
AcB/acb	3
aCb/acb	4

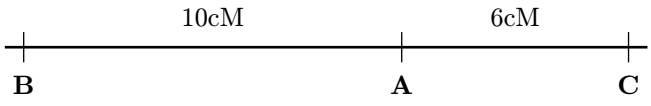
over these shorter distances, so we can now better estimate the map distance between **A** and **B** (which was our original goal) to be 37cM (instead of our original 23cM).

Note that this new experiment hasn't changed our estimate of the recombination frequency r_{AB} between **A** and **B**. It is still .23, since 23% of the gametes are recombinant with regard to these loci. We have just done a better job of estimating the true map distance between them. Notice that this process could continue if we wanted to be more accurate. In other words, we used the recombination frequency r_{AC} directly as an estimate of the map distance between **A** and **C**. But, maybe they are far enough apart that double crossovers may occur in that region. If we are worried about this, we may use another marker, say **D** between **A** and **C** to further analyze distances.

4.3.2 Ordering Loci and Map Construction

We will just discuss a simple case of constructing genetic maps considering only as many as three loci at a time. In this case, the process is simple. The problem we face is ordering the loci appropriately. Note that if we had already known the order, such as in the three-point testcross experiment above, this process would not be necessary. But in many cases we may have genetic map information from various individual experiments and so might not have prior knowledge of the ordering of loci.

First, we assume that our estimates of map distances between loci are accurate and therefore can be added. Let's say that we have three loci, **A**, **B**, and **C**, and have been given map distances from various experiments as $x_{AB} = 10$, $x_{BC} = 16$, and $x_{AC} = 6$. There are three possible orderings, depending on which locus should be in the middle (we won't be concerned with whether the overall direction of the ordering is correct). One possibility is **ABC**, another is **ACB**, and another is **BAC**. It is clear that the only one that makes sense, given the additivity of map units, is **BAC**. We would see that the correct genetic map can be pictured as



4.4 DNA Fingerprinting

The use of *DNA Fingerprinting* for identification of individuals has become very prevalent in recent years. Another term that can be used interchangeably is *DNA Profiling*. It is used in identifying and convicting criminals, proving suspected criminals to be innocent, in paternity cases to determine the father (or even mother, or both) of a child (e.g., the "Tsunami Baby 81" story), to identify plane crash victims, to identify long-awaited Vietnam war remains (as was recently in the news), to confirm the identify of terrorists, along with many other uses. It has served a very legitimate and useful purpose in society and will continue to do so.

The concept of DNA fingerprinting follows the same concept as for "real" fingerprints. Real fingerprints are used as an identification tool because it has been found that no two people have the same fingerprints (including identical twins), or at least as far as we've been able to detect after many years of scientific discovery on the topic.

DNA fingerprints are used for identification for the same reason, with one caveat. No one will claim that two people do not have the same DNA fingerprint. There are two reasons for this. One is that fact that it really depends on the detail of the DNA profile that is worked up on a person. If it is not very detailed, then there would in fact be a high likelihood, even a guarantee, that possibly many people have that DNA profile. Relate this to an attempt to identify someone uniquely based on eye color. It clearly can't be done because that trait does not provide enough information by itself. So, DNA profiles can range from very informative to less informative. Clearly, we will want to focus on the very informative types of profiles.

A second reason that it is not claimed that no two people have the same DNA profile is the following. Even if the profile is very informative, we will see that there is some probability (albeit, infinitesimally small) that can be attached to the event of two people having that profile.

Finally, it is in fact the case that identical twins have identical DNA, and so would have identical profiles.

4.4.1 Basics of a DNA Profile

We will focus our discussion on DNA fingerprinting for humans since this clearly has the most important sociological implications. We have seen earlier that if you compared the sequence of bases in the DNA of one person to another, you will find about 99% of it to be a perfect match. In other words, most of the 3 billion bases pairs that make up the DNA of an individual are exactly the same from one person to the next. Clearly, basing a DNA profile on these areas of DNA would have no use in unique identification.

The interesting areas of DNA, then, are the other 1% of base pairs. Such regions of DNA where differences do exist between individuals are called *polymorphisms*. This is not new news to us. All of our discussions and examples regarding genes and markers in DNA to this point have been predicated on the fact that different people have different DNA sequences in those regions. They would have had no interest to us if this weren't the case.

The types of polymorphisms we will be interested in are not related to genes, but to what we have called markers (non-coding regions of DNA). One example of a type of artificial marker that has been useful in DNA analysis is restriction fragment sites (also called restriction fragment length polymorphisms or RFLPs), which we have already discussed. As we have seen, these are regions of DNA where the differences between the DNA of individuals is detected by observing the different length of fragments left over after digesting that region with a restriction enzyme. Other common examples of such polymorphic regions that have been detected in humans are (there are others as well)

- *SNPs*, or *Single Nucleotide Polymorphisms*. These are regions of DNA which are the same from one person to the next, except for a potential differences in the base at one particular location. So, for example, for a given region of DNA, it may have been discovered that the following two DNA sequences are the only differences that exist from one person to the next:

  GCCATTG
  GCCGTTG

  The only difference is whether the base in the 4th position of this sequence is an A or a G. This is the most detailed type of polymorphism discovered to date and is a fairly recent finding due to the extreme difficulty of analyzing DNA at such a fine level.

  Notice that for a particular SNP marker, there would in theory be four possible different alleles in the population (one for each base that could be in that polymorphic position), although in practice SNP markers have been found to only have two variants as in the example above.

- *STRs*, or *Short Tandem Repeats* (also called *microsatellite markers*). These are regions of DNA that have been found where a certain base sequence repeats itself over and over, but for a different number of times in different individuals. For example, say for a particular STR marker region, the repeat sequence is GCC. It may be that some individuals have 3 repeats of GCC at that location, others have 5 repeats, while others have 33 repeats and other numbers in between. Such loci have been found to be highly polymorphic, meaning that there exist many possible versions (i.e., many different alleles) in the population.

  Another variation on STRs is VNTRs (*Variable Number of Tandem Repeats*). STR is usually used to refer to regions where the repeat sequence is fairly short, say 3 or 4 bases, where a VNTR locus is one where the repeat sequence is longer, maybe 10 to 15 bases. The existence of VNTRs were, naturally, found before STRs.

Many thousands of such markers have been found to exist across all chromosomes of the human genome. Very detailed genetic and physical maps of their locations have been constructed. Please see this link for a good review of the history of constructing human genetic maps using various markers, and this 1994 article in Science regarding the first very detailed (1cM resolution) genetic map, primarily constructed using STR

markers.

One typical fact about such DNA markers that makes them useful for DNA profiling is that they tend to each have many different alleles present in the population. That is not true of many genes. For example, the ABO blood group gene has three alleles and disease genes typically have two alleles as we have discussed. As we have seen above, markers have 4, 5, 10, even 20 or 30 alleles present in the population, and some have been found with hundreds.

How is this useful for DNA profiling? Let's start with a simple case where we are using one STR marker that has 15 alleles (let's call the marker **A** and its alleles A_1 through A_{15}). With 15 alleles present in the population, there are $\frac{15\times16}{2} = 120$ possible genotypes in the population, namely A_1A_1, A_1A_2, $\cdots$, A_1A_{15}. This is a fairly large number of different categories into which a person can be classified. As long as each of those categories is relatively infrequent in the population, this one marker is a good start to creating a useful DNA profile.

The downfall, clearly, is that while 120 is a large number of classifications, it by no means provides us method for uniquely classifying individuals. All things being equal, an average of 50 million or more human beings would fall into each of those categories.

To improve the detail of our DNA profile, we must consider a number of such loci at once. Let's say that we have six such markers that we will use to develop a DNA profile. Let Table 4.16 describe the notation and information about these markers.

Table 4.16: Details about a sample 6-locus DNA profile

Marker	Number of Alleles	Allele Notation	Number of Genotypes
A	15	$A_1, \cdots, A_{15}$	120
B	20	$B_1, \cdots, B_{20}$	210
C	10	$C_1, \cdots, C_{10}$	55
D	20	$D_1, \cdots, D_{20}$	210
E	15	$E_1, \cdots, E_{15}$	120
F	12	$F_1, \cdots, F_{12}$	78

Now, an example of a profile that a person may have across these six markers is $A_{10}A_{12}B_3B_8C_2C_{14}$ $D_5D_5E_1E_{10}F_{11}F_{12}$. The number of different profiles that exist in the population can be seen to be the product of the values in the last column of the above table. This is the number 2,724,321,600,000 - over 2.7 trillion! With about 6.5 billion people in the world, just six markers provides us a method that can serve to, for all intensive purposes, uniquely identify a person. We will see some other issues that must be raised in the use of such profiles.

4.4.2 DNA Profile Probabilities

We will now discuss the next important step in using a DNA profile, namely, the attaching of a probability to a particular profile. We will do this with respect to the use of DNA fingerprints in criminal investigations and trials.

Forensic DNA Fingerprints

When a crime has been committed, there are often a number of pieces of evidence left behind or otherwise known about the crime. One type of evidence that interests us in this discussion is biological evidence from which DNA can be extracted. For example, this can be in the form of hair, blood, saliva, skin cells, or semen in the case of a sexual crime. When investigators find such evidence, they can work up a standard DNA profile on it. As an example, let's say that a blood stain has been found at the scene of a crime, and the DNA profile found from that stain is (using the same example of having six markers that we discussed above):

$$A_{10}A_{12}B_3B_8C_2C_{14}\ D_5D_5E_1E_{10}F_{11}F_{12}$$

Clearly, this could be the victim's blood, so that must first be ruled out. This can easily be done by taking a DNA sample from the known victim (whether dead or alive), profiling it, and comparing to above. Let's say that for this particular blood stain, there is no match to the victim. We then know it is the blood of someone else who we presume must have been at the scene of the crime.

Now, let's say that through other avenues of the criminal investigation, there are two suspects being held with respect to this crime. With their consent, investigators can get a DNA sample from each and create a profile. The two profiles are

$$\begin{aligned} \text{SUSPECT 1} &\longrightarrow \quad A_{10}A_{12}B_3B_8C_2C_{14}\ D_5D_5E_1E_{10}F_{11}F_{12} \\ \text{SUSPECT 2} &\longrightarrow \quad A_2A_6B_4B_9C_3C_3\ D_1D_8E_6E_{14}F_7F_{10} \end{aligned}$$

A quick analysis shows us that Suspect 2 is not the one who left the blood stain in question at the scene. His profile does not match, so clearly it was not his blood. Note that does not necessarily mean that he did not commit the crime or wasn't otherwise involved, but that particular piece of evidence cannot be used against him.

More interesting for our purposes is the profile of Suspect 1. It does match the crime scene stain profile. Does it mean that it was his blood? Our discussion above seems to give some strong indication that it is, but we need to be more quantitative in our analysis. The question that is raised is: "What is the probability that a randomly selected person from the population has this profile?" If that probability is somewhat high, it takes some of the pressure off of this suspect because there are potentially a few or possibly many others who could have left it. If that probability of quite low, it begins to cast strong evidence that it must be that suspect's blood.

Such numerical calculations of strength of evidence are often presented in a courtroom setting by the use of a *likelihood ratio* (LR) (also called an *odds ratio*). This ratio is

$$\text{LR} = \frac{\text{P(This evidence was left at crime scene } \mid \text{ Suspect is guilty)}}{\text{P(This evidence was left at crime scene } \mid \text{ Suspect is innocent)}}$$

This ratio can be interpreted as how much more probable the evidence is if the suspect left the crime scene stain than if some random unknown person did (Evett and Weir, 1998). A higher value suggests more evidence of guilt, while a lower value suggests innocence.

The numerator calculates the probability that this DNA evidence would have been found conditional on the suspect being the guilty party. It should be clear that the numerator is, then, exactly one. If the suspect is guilty, then he must have left the stain and so that's why we got the match.

The denominator calculates the probability that this DNA evidence would have been found if it did not come from the suspect. In other words, in the population at large, what is the probability of this profile?

With these interpretations of the numerator and denominator in mind, we can write this likelihood ratio as

$$\text{LR} = \frac{1}{\text{P(Random person has this DNA profile)}}$$

So the calculation comes down to the probability in the denominator, and we will discuss some models for helping us make this calculation.

Datasets

Before discussing calculation methods, we should first understand the available datasets that have been collected from individuals. In particular, we will continue to discuss DNA profiles with relation to crime scene investigations. The FBI has collected blood samples from many individuals of different ethnic backgrounds, profiled them, and placed the information in databases. Early such profiles were based on VNTR loci, but more recently have been based on STR loci. An example of data collected from early VNTR profiles at six loci is shown in Table 4.17, and sample data for the D1S7 locus in the FBI's black database is shown in Table 4.18. See Budowle (1991) for details on the loci themselves.

Notice that Table 4.18 is large, and somewhat sparse. In other words, the sample size is somewhat small compared to the number of genotypes. There are many zeroes in the table, and most counts are on the small side.

Simple Multinomial Model

Now consider what the data table would look like if we combined the information from all six loci into one DNA profile. It would be impossible to actually show because it would have trillions of categories (210,673,710,000,000 = 210 trillion+ to be exact in the case of the FBI Black database). And since the

Table 4.17: Summary statistics for the FBI databases for six VNTR loci.

Locus	Race	Sample Size	No. of Alleles	No. Possible Genotypes	No. Observed Genotypes
D1S7	Caucasian	593	26	351	217
	Black	359	26	351	190
	Hispanic	520	24	300	197
D2S44	Caucasian	790	21	231	166
	Black	474	24	300	180
	Hispanic	514	19	190	144
D4S139	Caucasian	593	17	153	112
	Black	446	18	171	131
	Hispanic	521	16	136	105
D10S28	Caucasian	428	23	276	180
	Black	287	24	300	172
	Hispanic	439	21	231	165
D14S13	Caucasian	750	24	300	169
	Black	523	25	325	201
	Hispanic	493	23	276	198
D17S79	Caucasian	775	13	91	47
	Black	549	15	120	78
	Hispanic	521	9	45	39

sample size is only 359, the count for almost all of them would be 0, with 359 1's spread around thinly. There is some extremely small chance of a 2 showing up in such a table.

Despite the size of this table, the model that describes our data is multinomial, as we have come across many times before. In other words, we have a random sample of individuals and are classifying each into one of k categories. It just so happens that in this case $k \approx 210$ trillion.

That fact itself doesn't have to deter us. We know how to estimate probabilities from multinomial models using maximum likelihood. Our model is

$$(X_1, \ldots, X_{210 \text{ trillion}}) \sim \text{Multinomial } (n, P_1, \ldots, P_{210 \text{ trillion}})$$

and our maximum likelihood estimates of the unknown P's are

$$\hat{P}_i = \frac{X_i}{n}$$

Table 4.18: Genotypic counts for the FBI Black database at locus D1S7. The bottom row shows the observed counts for the 26 alleles. The sample size is $n = 359$.

1	0																								
2	0	0																							
3	1	0	1																						
4	0	0	0	0																					
5	0	0	1	0	0																				
6	0	0	0	0	0	0																			
7	0	0	0	0	0	0	0																		
8	1	0	0	0	1	0	0	0																	
9	0	0	0	1	0	0	0	1	0																
10	0	0	0	0	2	0	0	1	0	0															
11	0	0	0	1	0	0	1	1	0	1	2														
12	0	0	0	0	1	1	0	0	1	0	0	1													
13	2	0	0	0	1	1	1	0	1	1	1	1	2												
14	1	1	0	0	1	0	0	4	1	1	1	2	0	3											
15	0	0	0	1	0	3	0	1	2	4	3	2	3	3	2										
16	0	0	1	1	1	2	2	2	2	1	2	0	1	5	2	4									
17	0	0	1	0	0	0	0	2	1	0	0	0	1	2	1	1	3								
18	0	0	0	0	1	1	0	0	1	0	1	0	3	2	1	0	3	0							
19	0	2	1	0	0	0	0	2	2	2	0	1	1	1	1	2	2	5	2						
20	0	1	0	0	1	1	1	0	2	3	0	0	2	1	2	3	4	4	1	4					
21	0	1	1	0	0	3	0	2	0	0	1	1	2	1	1	3	3	2	1	2	6				
22	0	0	0	0	1	0	1	2	0	2	0	1	2	2	8	1	2	2	0	2	3	2			
23	0	1	0	1	0	1	0	2	0	1	3	2	2	4	4	6	3	4	2	4	2	3	1		
24	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	2	1	0	0	0	0	2	
25	0	0	0	0	0	0	1	0	0	1	1	1	1	1	1	1	1	2	3	2	1	1	2	0	0
26	0	1	0	0	1	2	1	2	1	1	0	2	1	4	0	6	2	2	1	3	4	6	4	2	0	3
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
	5	7	8	5	12	15	8	25	16	21	21	18	32	45	47	53	36	36	35	47	46	43	53	12	20	52

However, there is a problem we will have with the usefulness of such estimates. As we saw above, with such datasets X_i will typically always be either one or zero for each category. So some of our genotype estimates will be $1/n$ (in the cases where there had been an individual in the sample for that profile) and some (most) will be zero.

For the situations where $1/n$ is the estimate, $1/n$ itself is not such an unusually small number. For the FBI database above, $1/n = 1/359$. This is small, but with regard to a DNA profile and evidence to be used in court, it is not unusually small. It still leaves the chance that millions of people potentially share that profile, which is not very convincing courtroom evidence.

For the situations where zero is the estimate, this is also not useful. Zero is certainly a small number as far as probabilities go (the smallest, in fact), but it suggests that no one has that profile. In fact we know that someone does have that profile, namely our suspect. So the estimate zero is flawed also.

Generally, what is going on is that when we have a multinomial model, but the data is very sparse compared to the number of categories, the maximum likelihood estimators are not necessarily helpful. We just have not collected enough data to make convincing estimates of so many category probabilities.

Independence Within and Between Loci

In order to make better estimates of the probability of a DNA profile, we'll have to make further assumptions about our model and our data. The main assumption that we'll make use of is independence. As we'll see,

this will allow us to think of the profile as many individual and independent pieces of evidence instead of just one piece of evidence.

First, let's do a quick review of independence from probability theory. We say that two events (call them A and B) are *independent* if $\mathrm{P}(A \text{ and } B) = \mathrm{P}(A) \times \mathrm{P}(B)$. In other words, if the probability of both of them happening at the same time equals the product of the probabilities of them happening individually, they are independent events (of course, since this is a definition, the implication works in both directions).

As a simple example related to courtroom evidence, consider the following situation. Let's say that a crime occurred and a witness made two statements about the perpetrator: The perpetrator was a male and he had blue eyes. These are two pieces of evidence and we can consider them as an evidentiary profile of the criminal. If we have a suspect in hand who is male and has blue eyes, we might now be interested in calculating the probability of such evidence. As with our DNA profile example, this would be a calculation of the probability of this evidence matching any random person in the population.

This evidence is actually a combination of two pieces of evidence: information about the gender of the perpetrator and the eye color of the perpetrator. For this example, we could easily argue that these are actually two independent pieces of evidence. In other words, in the population, the traits of gender and eye color don't tend to go hand-in-hand with each other. So, our calculation of the probability of such evidence would be

$$\mathrm{P}(\text{Male and Blue Eyes}) = \mathrm{P}(\text{Male})\ \mathrm{P}(\text{Blue Eyes})$$

according to the independence rule.

Notice that if the evidence had been that the individual was male and drives a Harley motorcycle, we wouldn't have been able to make such a calculation. This is because we wouldn't consider these to be two independent pieces of evidence. It is well-understood that most people who drive Harleys are male, and so if you know one of these facts, the other is not anything new. So

$$\mathrm{P}(\text{Male and Drives Harley}) \neq \mathrm{P}(\text{Male})\ \mathrm{P}(\text{Drives Harley})$$

Now, back to our DNA profile evidence. We had wanted to calculate

$$\mathrm{P}(A_{10}A_{12}B_3B_8C_2C_{14}\ D_5D_5E_1E_{10}F_{11}F_{12}).$$

and found that the simple estimates from a multinomial model won't help us that much. However, if it can be stated that the evidence from each of the markers are independent of each other, we can instead calculate

$$\mathrm{P}(A_{10}A_{12})\mathrm{P}(B_3B_8)\mathrm{P}(C_2C_{14})\mathrm{P}(D_5D_5)\mathrm{P}(E_1E_{10})\mathrm{P}(F_{11}F_{12}).$$

Further, if we can also say that each marker itself is in Hardy-Weinberg Equilibrium (which is actually just a form of independence), we can further write the above as

$$2\text{P}(A_{10})\text{P}(A_{12})2\text{P}(B_3)\text{P}(B_8)2\text{P}(C_2)\text{P}(C_{14})\text{P}(D_5)\text{P}(D_5)2\text{P}(E_1)\text{P}(E_{10})2\text{P}(F_{11})\text{P}(F_{12})$$

Now, since each of these twelve probabilities will be somewhat small (typically on the order of 1/10 to 1/20), we can wind up with a very small number. If I substitute 1/15, for example, for each of these probabilities, we would calculate:

$$\text{P}(A_{10}A_{12}B_3B_8C_2C_{14}\ D_5D_5E_1E_{10}F_{11}F_{12}) = 0.00000000000007707 = 7.7 \times 10^{-14}$$

Now, this is evidence that could be useful in court.

We have already discussed tests for HWE. Of course, in these situations, the markers have very many alleles, so the models become much more complex and the calculations can only be done with the help of special computer programs. But otherwise, the theory of such tests is just as discussed before.

Once we have established that each marker individually is in HWE, we can proceed to test for independence between markers. For our purposes, it will suffice to just test for independence between pairs of markers instead of for all subsets of markers.

Let's setup the hypotheses we want to test. First, without loss of generality, we will refer to the two loci under consideration as **A** and **B**. Locus **A** has n_A alleles labelled $A_1, \cdots, A_{n_A}$, amd locus **B** has n_B alleles labelled $B_1, \cdots, B_{n_B}$. The total number of two-locus genotypes possible is $\frac{n_A(n_A+1)n_B(n_B+1)}{4}$, and so this is the number of multinomial categories in our full model (represented in the alternative hypothesis). The probability of a genotype will be notated by P with the genotype in the subscript (e.g., $P_{A_6A_9B_3B_5}$)

The null hypothesis states that the probability of any genotype can be written using the concept of HWE within a locus and independence between loci. In other words, we have the hypothesis

$$\text{H}_0 : P_{A_iA_jB_kB_l} = \begin{cases} 4P_{A_i}P_{A_j}P_{B_k}P_{B_l} & i \neq j; \quad k \neq l \\ 2P_{A_i}P_{A_j}P_{B_k}^2 & i \neq j; \quad k = l \\ 2P_{A_i}^2P_{B_k}P_{B_l} & i = j; \quad k \neq l \\ P_{A_i}^2P_{B_k}^2 & i = j; \quad k = l \end{cases}$$

for all combinations of i, j, k, l where $j \geq i$ and $l \geq k$.

Of course, this is a very complex null hypothesis. But, although messy, this just gives us a set of well-defined rules for writing genotype probabilities in terms of allele probabilities. Therefore, just like the test for HWE we saw previously, the null hypothesis can be used as the basis for computing expected counts for a goodness-of-fit test. We do have $n_A + n_B$ parameters to estimate in the null hypothesis (the allele frequencies for each locus), and this can be done using simple allele-counting MLE's as before. Then, our goodness-of-fit test would be setup as in Table 4.19.

Table 4.19: Setup of Goodness-of-fit Test for Independence Between Loci

Genotype	Observed Count	Expected Count
$A_1A_1B_1B_1$	X_{1111}	$n(\widehat{P_{A_1}})^2(\widehat{P_{B_1}})^2$
$A_1A_1B_1B_2$	X_{1112}	$2n(\widehat{P_{A_1}})^2(\widehat{P_{B_1}})(\widehat{P_{B_2}})$
$A_1A_1B_1B_3$	X_{1113}	$2n(\widehat{P_{A_1}})^2(\widehat{P_{B_1}})(\widehat{P_{B_3}})$
$\vdots$		
$A_6A_9B_3B_5$	X_{6935}	$4n(\widehat{P_{A_6}})(\widehat{P_{A_9}})(\widehat{P_{B_3}})(\widehat{P_{B_5}})$
$\vdots$		
$A_{n_A}A_{n_A}B_{n_B}B_{n_B}$	$X_{n_An_An_Bn_B}$	$n(\widehat{P_{A_{n_A}}})^2(\widehat{P_{B_{n_B}}})^2$

The table has $\frac{n_A(n_A+1)n_B(n_B+1)}{4}$ rows, and so the chi-square test statistic has that many terms added together (of the same form as before: $(O-E)^2/E$). Clearly, this is meant for computers to do the work.

The degrees of freedom for the test statistic will be $\frac{n_A(n_A+1)n_B(n_B+1)}{4} - 1 - (n_A - 1) - (n_B - 1)$. The problem that occurs is that with relatively small sample sizes (as we saw in the FBI dataset), the expected counts are not likely to meet our requirements for using the chi-square distribution. Fisher's exact test is the more reliable methodology, although it is computationally difficult.

Chapter 5

Sequence Analysis and Alignment

5.1 Overview

In this chapter, we will discuss methods for analysis of sequences of DNA or amino acids. As we have alluded to with a few examples earlier in the course, sometimes data comes to us as a DNA sequence (i.e., a sequence of A's, C's, G's, and T's) from a certain organism, or a sequence of amino acids that represents a protein synthesized by an organism.

There are many questions we might have about this kind of data. In fact, it is the enormous amount of sequence data becoming available that could be said to have pushed bioinformatics to the forefront in genomic analysis. In addition to the human genome project (which is now complete), there are many DNA sequences that have been typed and recorded across many species and research projects. Many (if not most) of this data is held in a publicly available and searchable database called GenBank (http://www.ncbi.nlm.nih.gov/Genbank/index.html). The front page of the site states the following

> "There are approximately 59,750,386,305 bases in 54,584,635 sequence records in the traditional GenBank divisions and 63,183,065,091 bases in 12,465,546 sequence records in the WGS division as of February 2006."

The amount of available data and the potential for analysis is enormous.

5.2 Single DNA Sequence Analysis

Our first class of analyses will have to do with analyzing a single DNA sequence. Such data can represent a gene or a non-coding region in an organism. For example, here is the DNA sequence from an Immunogloblin gene in mice downloaded from GenBank, consisting of 372 bases.

```
  1 atcttctttc tggtagcaac agctacaggt gtgcactccc aggtccagct gcagcagtct
 61 gggcctgagg tggtgaggcc tggggtctca gtgaagattt cctgcaaggg ttccggctac
```

```
121 acattcactg attatgctat gcactgggtg aagcagagtc atgcaaagag tctagagtgg
181 attggagtta ttagtactta caatggtaat acaaactaca accagaagtt taagggcaag
241 gccacaatga ctgtagacaa atcctccagc acagcctata tggaacttgc cagattgaca
301 tctgaggatt ctgccatcta ttactgtgca agatactatg gtaactactt tgactactgg
361 ggccaaggca cc
```

5.2.1 Composition of Bases

A first, simple question might be whether or not the four bases are equally represented in the DNA sequence. For the DNA sequence above, the relative proportions of the bases are

Base	Count	Percentage
A	101	.271
C	83	.223
G	94	.253
T	94	.253

This suggests that there are more A nucleotides and fewer C nucleotides than the 25% that would have been expected if there were an equal distribution. There is no particularly interesting statistical inferences to be made since these results and conclusions are specific to the DNA sequence at hand. We are not trying to use these results to make conclusions about other sequences.

Of course, if we consider the above sequence as just a random sequence from the mouse genome, we can use it as a basis for making inferences about all mouse DNA. Maybe, for example, we have a hypothesis that the C nucleotide is not as abundant in mice. We could setup a simple binomial probability model that counts the number of C's we come across in this random sequence. The model would be $X_C \sim$ Binomial (n, P_C). P_C represents the unknown proportion of nucleotides that are C's in the entire genome. From our previous binomial distribution theory, we know that $\hat{P_C} = X_C/n$ and an approximate 95% CI for P_C is given by $\hat{P_C} \pm 2\sqrt{\hat{P_C}(1-\hat{P_C})/n}$. For our example, we would compute the following CI

$$.223 \pm 2\sqrt{\frac{.223(1-.223)}{372}} = .223 \pm .043 = (.180, .266)$$

With this short sequence, we are unable to say that the overall proportion of C's in the mouse genome is not 0.25, since that value falls in our 95% CI.

A little more interesting from a biological standpoint is data like that reported in Weir (1996) from an external database. The DNA sequence from different regions of two closely linked human fetal globin genes were analyzed with regard to base composition. The different regions were the regions flanking the genes (just to the left is called the 5' flanking region, and just to the right is called the 3' flanking region), introns, exons, and the region between the genes. The data is reproduced here

Region	Length	A	C	G	T
5' Flanking	1000	0.33	0.23	0.22	0.22
3' Flanking	1000	0.29	0.15	0.26	0.30
Introns	1996	0.27	0.17	0.27	0.29
Exons	882	0.24	0.25	0.28	0.22
Between Genes	2487	0.32	0.19	0.18	0.31

One can begin to make statements about the richness of A nucleotides in flanking regions as compared to the synthesized part of the genes themselves (the exons). Since these regions are functionally different (the 5' flanking region, for example, contains promotor regions that assist in the transcription process), we might deduce some molecular hypotheses regarding the usefulness of the A nucleotide in such regions and then study it further biologically.

A model we might use for doing this analysis could be the following. We can focus on the number of A bases in the 5' flanking region and the number of A bases in the exons. We'll assume both to be binomial sampling situations, and to occur in what we'll assume to be independent regions of DNA. So our model for these counts is

$$X_{a,5'} \sim \text{Bin } (n_1, p_1)$$

and

$$X_{a,\text{exon}} \sim \text{Bin } (n_2, p_2)$$

Our hypothesis is that $p_1 > p_2$, or in other words, that there is an overabundance of A bases in 5' flanking regions (promoter regions) as compared to in the exons of genes. Specifically, our null and alternative hypotheses are

$$\text{H}_0 : p_1 = p_2 \quad \text{versus} \quad \text{H}_a : p_1 > p_2$$

or written differently

$$\text{H}_0 : p_1 - p_2 = 0 \quad \text{versus} \quad \text{H}_a : p_1 - p_2 > 0$$

We can test these hypotheses using the technique from our discussion of generic hypothesis tests where we hypothesize a parameter equals a constant versus greater than that constant (or less than, or not equal to it). In our general discussion of hypothesis tests in Chapter 2, we noted that we can find the MLE for the parameter and use the fact that MLE's have normal distributions for large sample sizes. This is what we'll make use of here.

All we have to do is notice that the parameter that we are testing for is $p_1 - p_2$. It just happens to be the difference of two other parameters, but we can still consider it as our parameter of interest. We just need to find the MLE for $p_1 - p_2$ and the variance of that MLE. This is not difficult, because we already know how to find MLEs for binomial distributions. We know that

$$\hat{p_1} = X_{a,5'}/n_1$$

and

$$\hat{p_2} = X_{a,\text{exon}}/n_2$$

are the individual MLE's for p_1 and p_2, and therefore the MLE for $p_1 - p_2$ is just $\hat{p_1} - \hat{p_2}$.

We also need the to calculate the variance of this estimator. By our rules of variances, and the fact that we already know the variance of binomial MLEs, we can calculate

$$\sigma^2_{\hat{p_1}-\hat{p_2}} = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

So, our test statistic is the following

$$T = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}}$$

and this statistic has a standard normal distribution if the null hypothesis is true.

We can use this to carry through our test based on the data from the human fetal globin genes above. Our estimates are $\hat{p_1} = .33$ and $\hat{p_2} = .24$ with $n_1 = 1000$ and $n_2 = 882$. Plugging these into the test statistic formula we derived above, we get $T = 4.35$. This results in a p-value of 0.0000 from the standard normal distribution, so we reject the null hypothesis and conclude that there does seem to be an overabundance of A nucleotides in the promoter regions as compared to exons.

Another interesting analysis involves the composition of bases in a sequence with regard to restriction fragment sites (RFLPs). For example, the enzyme *AfaI* recognizes the bases GTAC and breaks the DNA after the T. Consider a randomly selected strand of DNA that is 1000 bp long. Assume that the base present in each position of this randomly selected strand is completely random (i.e., 1/4 chance for each base), and is independent of the base in other positions (i.e., an A at one site does not make the chance of an A at the next site any more or less, for example).

Now suppose we let this enzyme digest this random strand. Let the discrete random variable X equal the number of sites recognized by the enzyme along this strand. Let's come up with a reasonable approximate probability model for X.

A way to think about the problem is to consider looking at the strand base by base, and for each position notice if it is the start of the GTAC sequence. We can do this for 997 positions, since starting with position 998 we will not have four following bases to look at. Since we are counting up the number of events that happen over a fixed amount of space (in this case 997 DNA bases), we can think about using the Poisson distribution as a reasonable model. So let's say $X \sim \text{Poisson}(\lambda)$. Now we need to determine a reasonable value for λ. A simple way to think about it is that at each position, we can consider there to be a 1/256

chance of seeing the bases GTAC in order starting at that position (since each has a 1/4 probability in a random DNA sequence). Since there are 997 positions we will look at, the expected number of GTAC sequences we'd expect over the entire length is 997/256. We can use this as our Poisson distribution parameter, giving us $X \sim \text{Poisson}(997/256 = 3.8945)$ as an approximate model. Under this model, we would calculate the probability of having no *AfaI* recognition sites as $\text{P}(X = 0) = \frac{e^{-3.8945}(3.8945)^0}{0!} = .0204$. In other words, in a random DNA sequence, it is very likely to have at least one GTAC subsequence.

Also, by pretty much the same argument, we can consider a Binomial model such as $X \sim \text{Binomial}(997, 1/256)$ which is approximately the same as the above Poisson model with such a large sample size. This would give $\text{P}(X = 0) = \binom{997}{0} (1/256)^0 (255/256)^{997} = .0202$.

By the way, one reason for these being approximate models is that we really aren't quite analyzing 997 independent positions. For example, if we find the sequence GTAC starting at position 36, say, then we know it cannot also start at positions 37, 38, or 39 since it would be certain that a G would not be in those spots. So, the more GTAC's we find, the fewer actual positions we are analyzing. But this difference has only a minor affect on this model.

5.2.2 Independence of Consecutive Bases

Another interesting biomolecular question has to do with neighboring bases in a DNA sequence. We have come across problems before where we make the assumption that the base present in one position is independent of that in the previous position. But is this actually true in real DNA sequences? We can answer this question by setting up an appropriate probability model.

Consider a DNA sequence of n bases. Now, let's look at the sequence two bases at a time. For example, in the above mouse DNA sequence of 372 bases, we will consider our data as the 371 pairs we come across. The start of that sequence (first 10 bases) was

```
atcttctttc
```

We will consider the data from this part of the sequence as the consisting of the nine pairs

```
at tc ct tt tc ct tt tt tc
```

Considering these pairs as data, we can again use a multinomial probability model. We have 16 categories (each of the possible ordered pairs of bases) each with a certain probability of occuring under our model. That is:

$$(X_{aa}, X_{ac}, X_{ag}, \ldots, X_{tt}) \sim \text{Multinomial } (n-1, P_{aa}, P_{ac}, P_{ag}, \ldots, P_{tt})$$

The null hypothesis we might setup would be that these probabilities of ordered pairs are actually just the product of the relevant individual bases. In other words, we are hypothesizing

$$\mathrm{H}_0 : P_{ij} = P_i \; P_j \; ; \quad i, j \in \{a, c, g, t\}$$

This states a hypothesis of independence. That is, if consecutive bases really are independent of one another, then the probability of a certain pair should just be the product of the individual probabilities (which is the basic definition of independence).

So, our probability model under the null hypothesis is

$$(X_{aa}, X_{ac}, X_{ag}, \ldots, X_{tt}) \sim \text{Multinomial } (n-1, \; P_aP_a, \; P_aP_c, \; P_aP_g, \ldots, \; P_tP_t)$$

We can conduct this as a chi-square goodness-of-fit test with 16 categories. There are three unknown parameters (the individual base frequencies P_a, P_c, P_g, P_t) which need to be estimated using the standard sample proportion MLE estimates. For example, there are 101 A bases in the sequence, leading to the estimate $\hat{P}_a = 101/372 = .2715$, and so the expected count under H_0 for the aa category is $371 \times (.2715)^2 = 27.3$. The full goodness-of-fit table looks like

Pair	Observed Count	Expected Count
aa	21	27.3
ac	26	22.5
ag	32	25.5
at	22	25.5
ca	34	22.5
cc	16	18.5
cg	1	20.9
ct	31	20.9
ga	21	25.5
gc	23	20.9
gg	29	23.7
gt	21	23.7
ta	24	25.5
tc	18	20.9
tg	32	23.7
tt	20	23.7

The chi-square test statistic is $T = 40.65$ with 12 (16-1-3) degrees of freedom. The resulting p-value is 0.0000 leading us to reject the null hypothesis. This suggests we have strong evidence that there is not independence from one base to the next in mouse genes. In particular, notice the major lack of CG pairs (only 1 observed) in this data. This has been noticed in other genes as well (see Weir, 1996) and a biochemical reason has been suggested which we'll come across again later in the course.

This analysis can be extended to considering three or more bases at a time, with a substantial increase in the number of categories that need to be considered (e.g., for testing independence across three bases, we

would have $4^3 = 64$ multinomial categories in the model).

5.3 Comparing and Aligning Two Sequences

In this section, we will discuss methods for comparing two sequences to each other in order to make statements about how similar or dissimilar they are. In addition, for sequences that have some similarity, we will attempt to align them in order to analyze their subregions that are most similar. The sequences we will analyze will be either DNA sequences of bases, or protein sequences of amino acids.

As we have discussed, one of the main reasons a researcher would want to make such comparisons is the following. If he or she is attempting to better understand the function of a particular protein (produced from a gene) in a certain species, it would be helpful to know if a similar protein has already been studied, whether in that species or another. If the function and molecular structure of the protein is already well understood, it will help the researcher better understand what he or she is working with.

As a concrete example, suppose a human disease researcher has found a gene (and related protein) that is believed to be related to a certain disorder. Helpful research would focus on better understanding the function (or disfunction) of that protein in the body. It may turn out that a protein with a very similar molecular makeup has been found in the mouse genome and studied extensively. Maybe it had been found that this mouse gene/protein tends to act as an inhibitor, perhaps blocking certain antigens from attaching to blood cells. Now, although the protein that the researcher found in humans does not have exactly the same makeup, this is a starting point for better understanding of its function in the body, and ultimately how it is related to this human disorder. Although humans and mice may seem very different, there are in fact many highly conserved biological functions that are the same in both, and are carried out by very similar proteins.

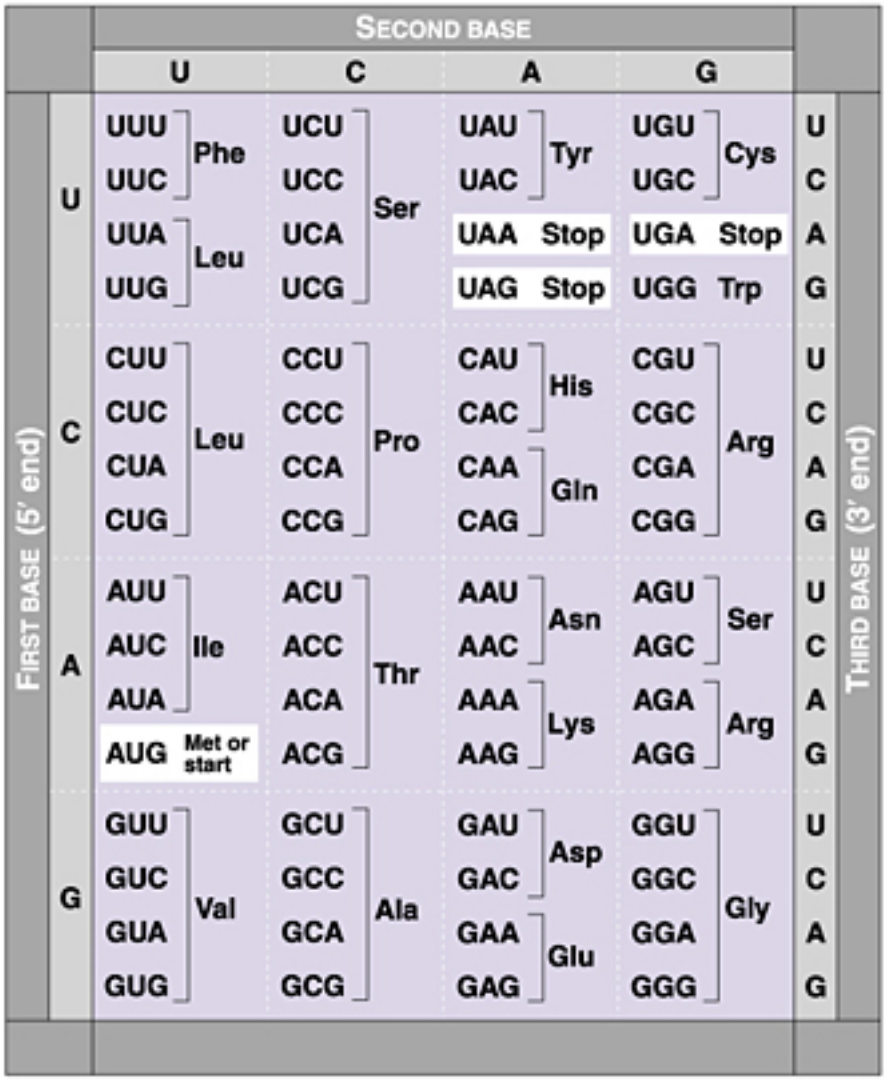
5.3.1 Evolutionary Background

Some knowledge of basic genetic evolutionary forces is necessary to better understand these procedures and their importance. We will discuss these more fully in lecture. The genetic code we saw in Chapter 1 is also repeated here since we will refer to it often.

5.3.2 Sample Sequences

Let's give two quick examples of real sequences that we might want to compare and/or align. First, we might have the following two DNA sequences

```
accctgacgaacttcgaaagct
accgtgacgaaattcgaaaaagct
```



Upon a quick inspection, we would notice that there seems to be some close similarity between these sequences. They are not exactly the same, and in fact one is longer than the other, but nonetheless, they are quite similar. The differences we notice are that there seems to be a couple of base substitutions (changes) from one to the other, and there seems to be what looks like an addition of bases in the bottom sequence. More specifically, there seems to be two extra A's toward the end of the bottom sequence that keep the sequences from aligning even more perfectly.

We will also compare protein sequences. This is probably the more typical sequence comparison situation. Recall that proteins are sequences of amino acids. When we write down a protein sequence, we use a standard one-letter (or three-letter) code for each amino acid. The one and three letter codes are shown in the following table.

Amino Acid	Three-letter	One-letter
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamine	Gln	Q
Glutamic acid	Glu	E
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

A sample amino acid sequence from a real human gene is below (GenBank reference). This is only the first 60 of the nearly 600 amino acids in this protein.

```
mapgqlalfs vsdktglvef arnltalgln lvasggtaka lrdaglavrd vseltgfpem
```

5.3.3 Statistical Significance of Matching Sequences

First, we might be interested in an understanding of how significant it is to see matches in a sequence. In other words, just because we notice some matching bases in a DNA sequence, or even a lot of matching bases, doesn't necessarily mean that is a significant event. We expect some matching, and possibly some long runs of matches, even for two completely random and independent sequences.

Let's consider two DNA sequences of lengths n_1 and n_2. Notice that we allow for them to possibly be of different lengths. Also, let the frequency of the four bases A,C,G,T in the first sequence be P_{A1}, P_{C1}, P_{G1}, and P_{T1}, and the frequencies in the second sequence are P_{A2}, P_{C2}, P_{G2}, and P_{T2}.

Say that we are analyzing the two sequences for a run of exact base matches of length r. Notice that there are $(n_1 - r + 1)(n_2 - r + 1)$ combinations of positions at which we can begin this search. For any particular such starting position, we might want to calculate the probability that a run of length r begins at that point.

The probability that the next r bases match, but the $r+1$(st) doesn't follows a geometric probability model. The probability of a match at each position is

$$P = P_{A1}P_{A2} + P_{C1}P_{C2} + P_{G1}P_{G2} + P_{T1}P_{T2}$$

In the geometric model, we count the number of trials (i.e., bases) until we get our first success (i.e., non-match). So, if we let

$$X = \text{Base that first non-match occurs from a given starting position,}$$

we have that $X \sim$ Geometric $(1-P)$, and the probability that we will see r consecutive matches is $\mathrm{P}(X = r+1) = P^r\ (1-P)$ from the standard geometric pmf. Again, keep in mind that a "success" in this model is a non-match, which is why our success probability is $1-P$ and not P.

For sequences where the bases are equally likely, the formula for P above reduces to 0.25, and we can calculate the probability of say 3 consecutive matches from a particular starting position as $\mathrm{P}(X = 4) = (.25)^3\ (.75) = .012$. So, it is not likely, but if we analyze many starting positions, we are likely to find three consecutive matches somewhere along the way even for random sequences.

A little more interesting would be analysis of the random variable L=length of longest match found from all such possible combinations of starting positions. We can use this random variable to determine how significant it is that we notice a certain long run of matches if we align the sequences in a certain way. This random variable is not easy to define or write down a probability distribution for based on what we have learned in this course. But others have done this, and it turns out that we can write down the mean and variance of L as (Arratia and Waterman, 1985)

$$\mu_L = \frac{2\ln n}{\ln(1/P)}$$

$$\sigma_L^2 = \frac{1.645}{(\ln P)^2} + \frac{1}{12}$$

We can take the n in the formula for μ_L to be the average of the two lengths, but should only use this is the two sequences are about the same length. Note that these are the mean and variance of L under the null hypothesis that the sequences are not related and are just completely random.

From the above, we can calculate, for example (see Weir, 1996), that if the sequences are both of length 100 and the four bases are all considered equally likely in both, then $\mu_L = 6.64$. This says that if we evaluate all possible ways to align the sequences and over all of these take note of the longest run of matches we find, on average we would expect to find that longest run to be of length 6.64 bases.

The standard deviation of L for this example would be

$$\sigma_L = \sqrt{\frac{1.645}{(\ln .25)^2} + 1/12} = 0.9692.$$

We can use this to conduct a hypothesis test based on the longest match we have noticed in a pair of sequences we are analyzing. Our hypotheses would be

H_0 :The sequences are random, unrelated sequences

H_a :The sequences are related in some way

And we can use the general hypothesis testing theory from the end of Chapter 2 to say that the test statistic:

$$T = \frac{L - \mu_L}{\sigma_L}$$

has a standard normal distribution if the null hypothesis is true.

To follow through with the above calculations, say we align the hypothesized sequences and notice a longest match of $L = 11$. We could calculate the test statistic as

$$T = \frac{L - \mu_L}{\sigma_L} = \frac{11 - 6.64}{0.9692} = 4.50$$

which has a standard normal distribution. This gives a p-value of 0.0000, meaning that this was a significant finding, and these two sequences are likely not just random sequences and so are probably related.

5.3.4 Aligning Sequences

Our next major topic of discussion will be to discuss algorithms for aligning two sequences. The general idea is that if we are given two sequences, whether DNA or protein, we would like to consider all possible ways of aligning the residues of one to another. The alignment that provides the best overall amount of matching between the two will be the alignment we consider the most likely description of the evolutionary divergence between the two sequences.

To take into account the possibility of insertions and deletions of residues that may have occurred during the evolution of one sequence to the other, we must also allow for *gaps* in our alignments. In other words, sometimes a residue from one sequence is best lined up with a gap in the other in order to keep the alignment between other residues on track.

In rare cases, the best alignment will be obvious. For example, if two protein sequences are

```
HGKKVAAL
```

and

```
HGKKVLAF
```

then the best alignment is pretty clearly

```
HGKKVAAL
||||| |
HGKKVLAF
```

There are two amino acids that don't match up in this alignment, but we can see that any other shifting in one direction or the other of the alignment will totally throw off the main matches.

Based on this alignment, our conclusion about what may have happened through evolution is that the first A amino acid in the top sequence was mutated to an L and the L in the top mutated to an F somewhere along the way (or vice versa...evolution may have worked from the second sequence toward the first). Perhaps these substitutions occurred through the mutation of a base in the underlying DNA, or maybe a series of mutations of bases.

Some other examples are not necessarily as obvious, and may require us to include gaps in our alignment to best match up the sequences. For example, consider the DNA sequences (which are of different lengths)

```
AATCTGGCT
```

and

```
AAGTCTAGGGT
```

If we look closely, we might notice the following alignment seems reasonable.

```
AA-TCT-GGCT
|| ||| || |
AAGTCTAGGGT
```

Gaps have been inserted into the top sequence to help better align the two. The gaps are denoted by dashes. With the gaps, we can now see that the two sequences are actually quite similar and match up nicely except in three spots - two of which are explained by gaps, and one mismatch (the C with the G towards the end). From an evolution standpoint, if we consider the top sequence the older sequence, we might surmise that two insertion mutations occurred in the past, one that inserted a G base after the first two A's, and the other that inserted an A base between the T and G. Also, apparently a substitution mutation from C to G occurred toward the end of the sequence.

The above examples show that in some simple cases where we have short sequences that have some obvious similarities, the alignment process is not difficult and can be done by inspection. However, for more realistic situations, we need an algorithm to help us do the alignment.

There are various algorithms for aligning sequences, a couple of which we will talk about. However, they are all based on the same basic concept. The idea is that we can consider all the possible ways of aligning the two sequences. For each of these possible alignments, we can assign a score to it that represents how likely that alignment accurately reflects the process of evolution from one sequence to the next. An aligned pair of

residues will receive a positive score if they match, and a lower score, or a negative score if they don't match or if a gap is used at that point.

For example, take the two DNA sequences from above. The alignment we proposed above would certainly be a high-scoring alignment since there are mostly matches, along with one mismatch and two inserted gaps. Another possible alignment that we could have considered is the following

```
AATCT---GGCT
 |  |    || |
-AAGTCTAGGGT
```

While this alignment is not totally off-the-wall (it still has five matches in it), it is clearly not as reasonable as the one above. The problem is that this alignment requires too many events to have happened through evolution to explain it. We would have had four insertion/deletion mutations and three substitution mutations. These events are relatively rare, and so alignments that in general minimize the number of such events are preferred, and better reflect the reality of evolution. So, of these two alignments, the first would score high and the second would score low, leading us to choose the first.

Now, of course, the difficulty is that considering and scoring all possible ways of aligning two sequences, including the possibility of gaps, is an enormous task once the sequences get just mildly long. If we have two sequences both of about length n, then the total number of possible alignments that we would need to consider (including possible gaps) is on the order of (Durbin, et al. 1998)

$$\frac{2^{2n}}{\sqrt{n}}$$

which quickly becomes very large and computationally infeasible even with computers. The algorithms we will use are dynamic programming algorithms that essentially search through this entire space in a way that ignores unlikely alignments and focuses on the likely ones. They are guaranteed to find the highest scoring alignment, but searching through all possible alignments is still computationally expensive and impractical for long sequences. We will later mention other methods based on dynamic programming which will be more computationally feasible even for long sequences.

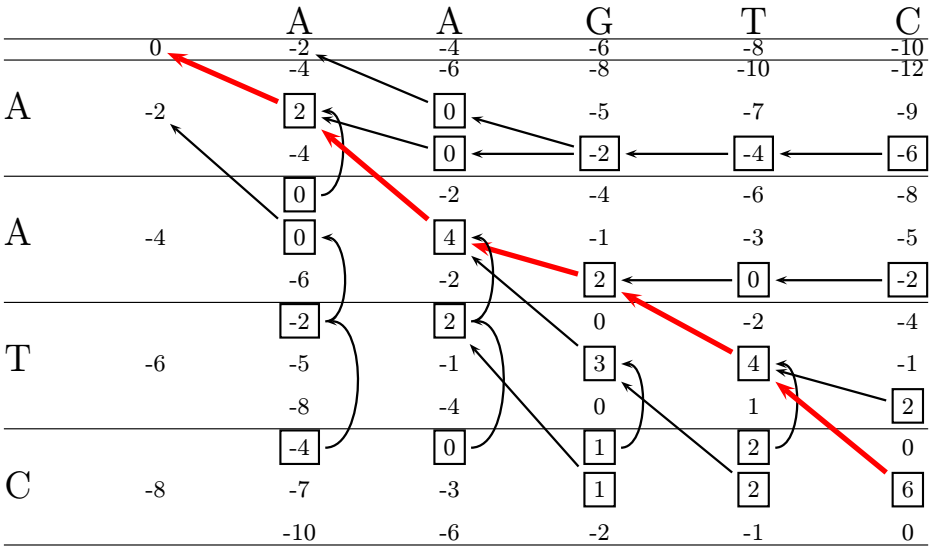
We will discuss the algorithms themselves in a bit, but the process of scoring can be discussed first since they are all based on the same concepts. It is a little simpler to discuss DNA alignment scoring systems, so let's focus on that first.

Typically, a simple scoring system is used. For example, we might score matching bases as +2, a mismatch as -1, and give a gap penalty of 2 (meaning a score of -2 for a base aligned with a gap). Or we may give a score of 0 for matches, -1 for mismatches, with a gap penalty of 3. Different scoring systems can be used, but all are generally on the same scale. In other words, scoring 15 for a match and -15 for a mismatch would not be a typical scoring system for DNA alignment.

Notice that for the first scoring system mentioned above, our first alignment of the two DNA sequences would receive a score of 11, and the second alignment would receive a score of -1.

The most popular algorithm for doing the scoring and performing the alignment is called the *Needleman-Wunsch* algorithm from their 1970 paper (Needleman and Wunsch, 1970), which was extended and made more efficient by Gotoh (1982). We will discuss this algorithm more fully in lecture.

Example: Align the DNA sequences aagtc and aatc using the Needleman-Wunsch algorithm. Score a match as +2, a mismatch as -1, and a gap as -2. Details will be discussed in lecture. Below is the resulting matrix and final alignment.



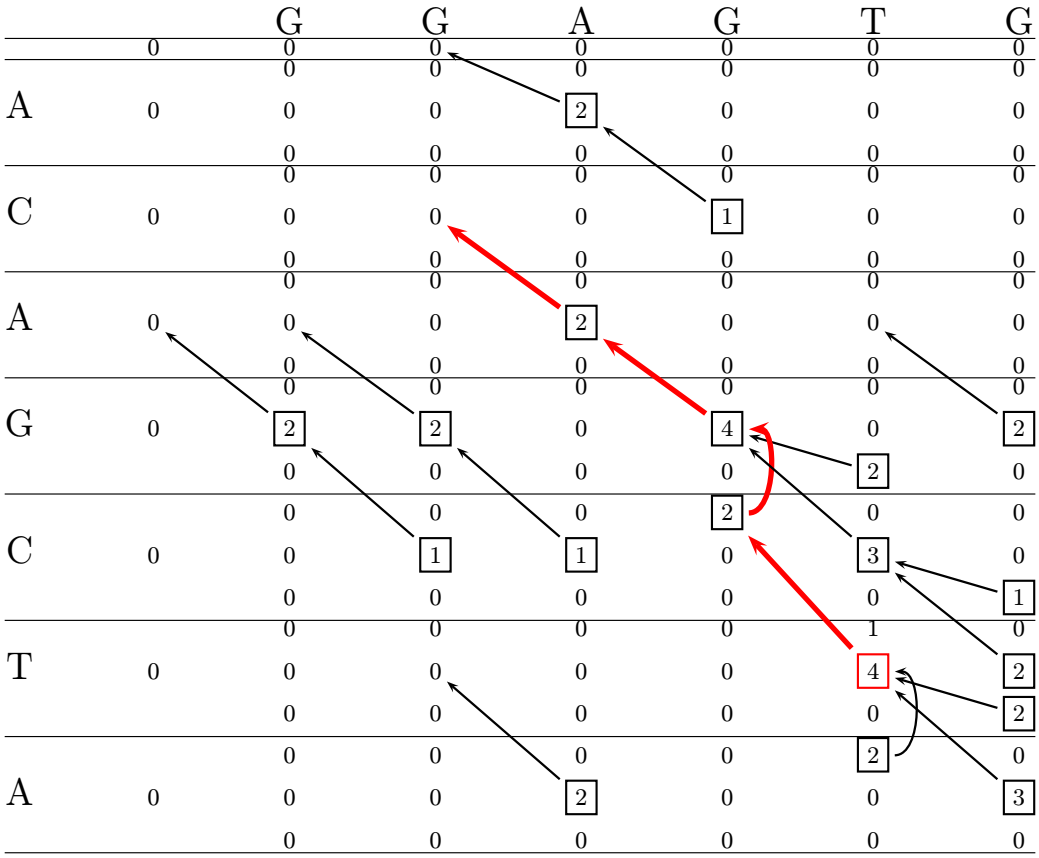
The traceback gives us the alignment

```
aagtc
aa-tc
```

with a score of 6.

A dynamic programming algorithm for *local alignment* of two sequences is called the *Smith-Waterman* algorithm. It is a simple extension of the Needleman-Wunsch algorithm and will be discussed more fully in lecture.

Example: Align the DNA sequences aagtc and aatc using the Smith-Waterman algorithm to perform a local alignment. Score a match as +2, a mismatch as -1, and a gap as -2. Details will be discussed in lecture. Below is the resulting matrix and final alignment.



The traceback gives us the local alignment

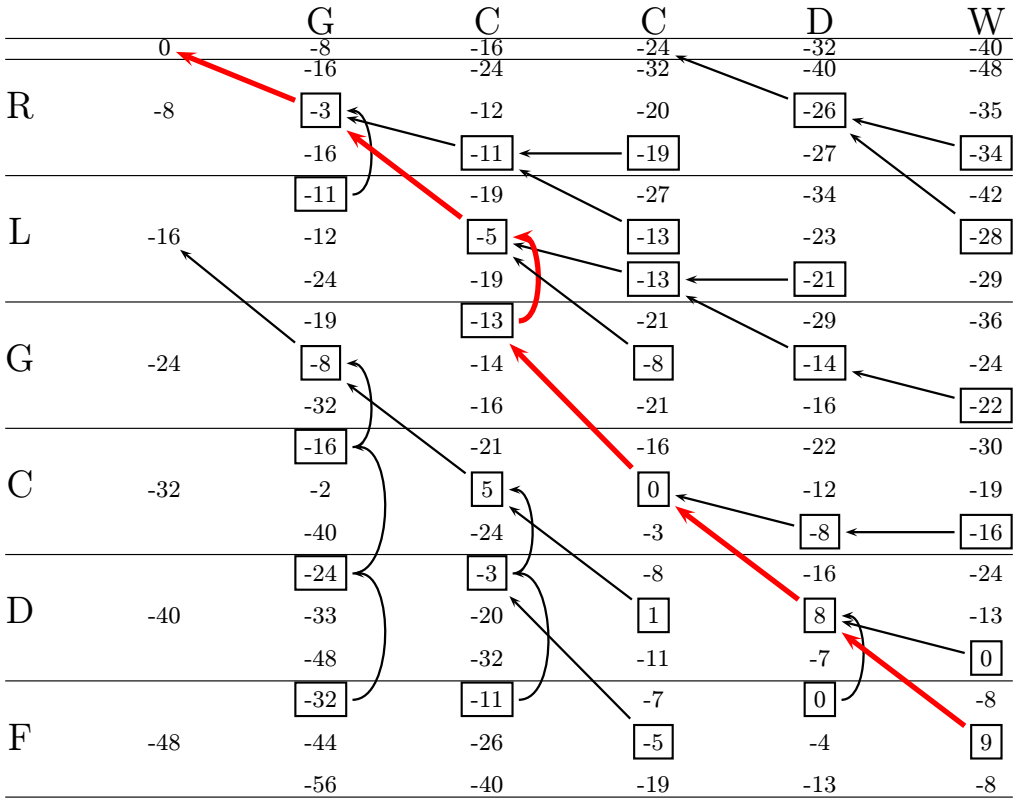
```
agct
ag-t
```

with a score of 4.

Example: Perform a global alignment of the protein sequences GCCDW and RLGCDF using the Needleman-Wunsch algorithm. Use the BLOSUM50 scoring matrix with a gap penalty of 8. Details will be discussed in lecture. Below is the resulting matrix and final alignment.

Figure 5.1: BLOSUM50 Scoring Matrix for Protein Alignment

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	**5**	−2	−1	−2	−1	−1	−1	0	−2	−1	−2	−1	−1	−3	−1	1	0	−3	−2	0
R	−2	**7**	−1	−2	−4	1	0	−3	0	−4	−3	3	−2	−3	−3	−1	−1	−3	−1	−3
N	−1	−1	**7**	2	−2	0	0	0	1	−3	−4	0	−2	−4	−2	1	0	−4	−2	−3
D	−2	−2	2	**8**	−4	0	2	−1	−1	−4	−4	−1	−4	−5	−1	0	−1	−5	−3	−4
C	−1	−4	−2	−4	**13**	−3	−3	−3	−3	−2	−2	−3	−2	−2	−4	−1	−1	−5	−3	−1
Q	−1	1	0	0	−3	**7**	2	−2	1	−3	−2	2	0	−4	−1	0	−1	−1	−1	−3
E	−1	0	0	2	−3	2	**6**	−3	0	−4	−3	1	−2	−3	−1	−1	−1	−3	−2	−3
G	0	−3	0	−1	−3	−2	−3	**8**	−2	−4	−4	−2	−3	−4	−2	0	−2	−3	−3	−4
H	−2	0	1	−1	−3	1	0	−2	**10**	−4	−3	0	−1	−1	−2	−1	−2	−3	2	−4
I	−1	−4	−3	−4	−2	−3	−4	−4	−4	**5**	2	−3	2	0	−3	−3	−1	−3	−1	4
L	−2	−3	−4	−4	−2	−2	−3	−4	−3	2	**5**	−3	3	1	−4	−3	−1	−2	−1	1
K	−1	3	0	−1	−3	2	1	−2	0	−3	−3	**6**	−2	−4	−1	0	−1	−3	−2	−3
M	−1	−2	−2	−4	−2	0	−2	−3	−1	2	3	−2	**7**	0	−3	−2	−1	−1	0	1
F	−3	−3	−4	−5	−2	−4	−3	−4	−1	0	1	−4	0	**8**	−4	−3	−2	1	4	−1
P	−1	−3	−2	−1	−4	−1	−1	−2	−2	−3	−4	−1	−3	−4	**10**	−1	−1	−4	−3	−3
S	1	−1	1	0	−1	0	−1	0	−1	−3	−3	0	−2	−3	−1	**5**	2	−4	−2	−2
T	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	2	**5**	−3	−2	0
W	−3	−3	−4	−5	−5	−1	−3	−3	−3	−3	−2	−3	−1	1	−4	−4	−3	**15**	2	−3
Y	−2	−1	−2	−3	−3	−1	−2	−3	2	−1	−1	−2	0	4	−3	−2	−2	2	**8**	−1
V	0	−3	−3	−4	−1	−3	−3	−4	−4	4	1	−3	1	−1	−3	−2	0	−3	−1	**5**



The traceback gives us the alignment

```
RLGCDF
GC-CDW
```

with a score of 9. Notice that another alignment that comes to mind that seems to make sense is

```
RLGC-DF
--GCCDW
```

This actually aligns more exact matches (3 instead of 2). But the score is only $-8 - 8 + 8 + 13 - 8 + 8 + 1 = 6$ which is why our dynamic programming algorithm didn't find it. The three gaps that need to be introduced are too much for the three matches to overcome. This is actually a good example of where a local alignment algorithm might work better because the initial gaps introduced at the beginning of the second sequence would not get penalized (the algorithm would start over from a score of zero when it aligns the G's). Try this using local alignment on your own.

Sometimes, we might want to use a more complex gap penalty. So far, we have used the same penalty for each gap introduced even if there are many consecutive gaps. But, it is often assumed that consecutive gaps are not independent of each other, and may represent one insertion or deletion evolutionary event. So, we might assign a smaller penalty for gaps in a series after the first gap. For example, we might use -8 as the penalty for the first in a gap series, but only -2 for the rest in that series. The -8 is often called the *gap-open* penalty (the penalty applied for opening, or starting, a gap), and the -2 is called the *gap-extension* penalty. Applying this to the example just above would lead us to calculating the following score:

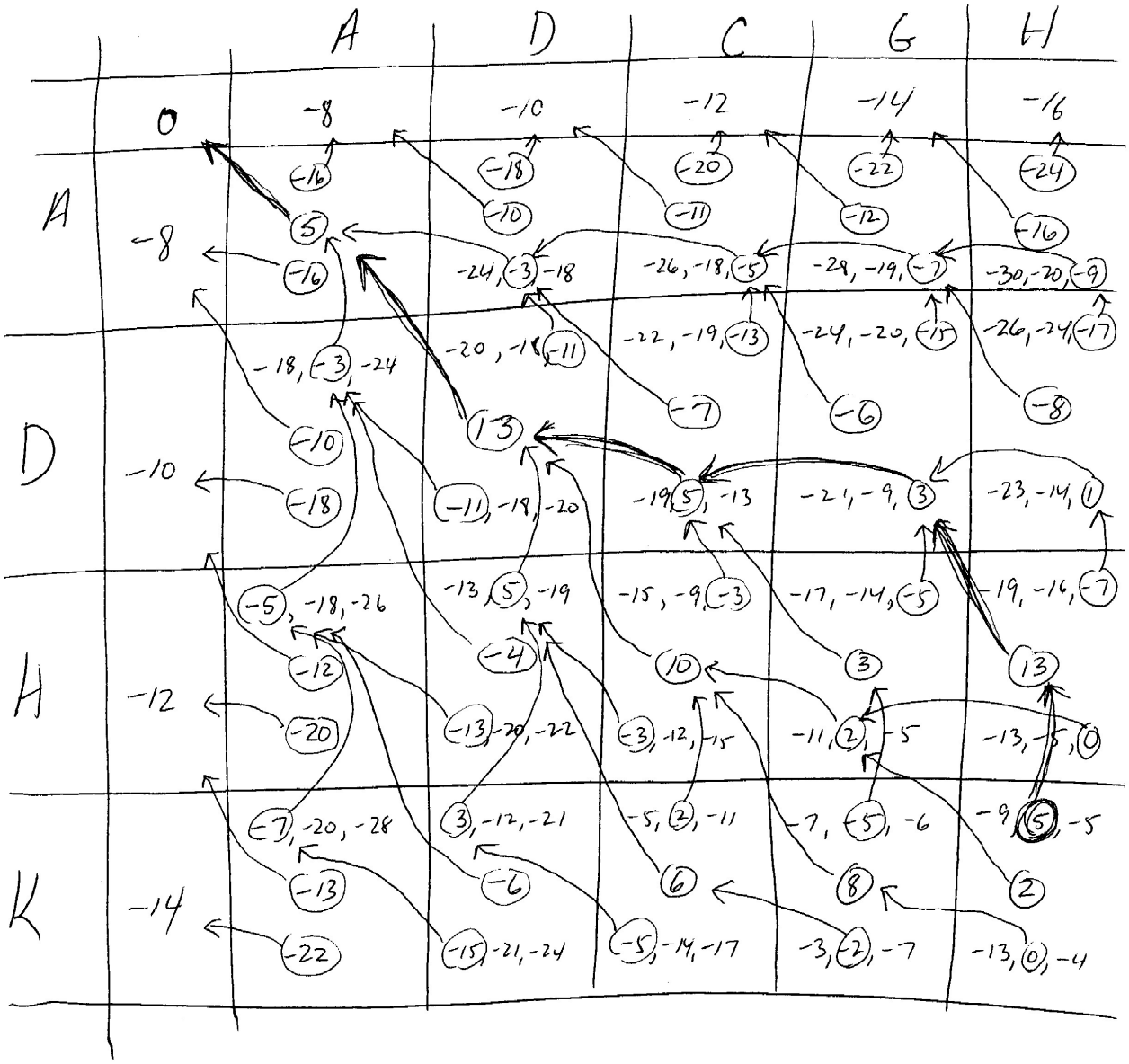
```
          R   L   G   C   -   D   F
          -   -   G   C   C   D   W
SCORE:  (-8)(-2)(8) (13)(-8)(8) (1)
```

This would give us a final score of 12, which would beat out the previous winning alignment. Less total penalty was assigned to the unmatched RL residues hanging off the left end on the top sequence.

Another common standard used in practice is to set the gap-open penalty to -12 and the gap-extension penalty to -2. Or -8 and -1 is sometimes used.

Actually performing the alignment with gap-extension penalties using a dynamic programming algorithm is similar to the Needleman-Wunsch-type algorithm. But, now we have a little more to keep track of throughout the process, which will be shown through example.

Example: Perform a global alignment of the protein sequences ADHK and ADCGH using the Needleman-Wunsch-type algorithm. Use the BLOSUM50 scoring matrix with a gap-open penalty of -8 and a gap-extension penalty of -2. Details will be discussed in lecture. Below is the resulting matrix and final alignment (sorry for the hand drawing, but there are too many arrows to deal with in LaTeX).



FINAL ALIGNMENT

A D – – H K

A D C G H –

Score = 5

NOTE: The arrow from the circled 2 in the lower right corner should go back to the circled 3 in the cell diagonally left and above it (not to the circled 2 in that cell).

5.3.5 Heuristic Algorithms: BLAST and FASTA

We have mentioned that dynamic programming algorithms like Needleman-Wunsch and Smith-Waterman are optimal, in the sense that they are guaranteed to find the highest scoring alignment. However, this property has a drawback, namely, speed. In a typical sequence alignment problem, one has access to a single sequence or a set of sequences, and wants to find alignments of these sequences to all other known sequences (for example, all sequences stored in GenBank) in order to find the one with which it aligns best. This is a big problem, as there are over 2.5 million protein sequences, and many more DNA sequences currently stored.

Faster algorithms, generally referred to as heuristic algorithms, have been developed that address the problem of computational time. The most widely used of these are *BLAST (Basic Local Alignment Search Tool)* and *FASTA*. They can be up to 50 times faster than dynamic programming algorithms (Mount, 2004). As always there is a tradeoff: these algorithms are not guaranteed to find the highest scoring alignment. However, studies have shown that they are indeed quite accurate, and typically will find the best, or near-best, alignment.

We will not discuss the details of how these algorithms work, but just give a quick overview. They are built on computer science string matching theory. The idea behind the algorithms is that the best alignment between sequences is likely to have at least some small region of perfectly matching residues in it. They make use of this idea by first searching for these small regions of perfect matching, and then focusing on just those regions (call them *seed* regions). After finding such a seed region, the search extends outward from there, essentially looking for a high-scoring local alignment in the neighborhood of that seed. After analyzing all such seed regions, the best alignments can be ranked and reported.

You can probably see why this does not guarantee finding the overall best local alignment. For example, it may be that the best alignment is one that contains a number of close substitutions and some gaps, and the regions of direct matches may never reach significant lengths. In any case, the algorithms work quite well and are capable of searching large databases very quickly.

The main server for BLAST can be found at `http://www.ncbi.nlm.nih.gov/blast/` (click "Getting started" on the left for a better overview of BLAST and the many options available). This server receives tens of thousands of search requests every day (Mount, 2004), and is a major hub of bioinformatics activity. The main server for FASTA is at `http://fasta.bioch.virginia.edu/`. You should become familiar with the options available for running BLAST searches, and try a few of your own.

5.3.6 Probabilistic Background for these Scoring Systems

So far, we have discussed common algorithms and calculations for aligning two sequences, but not the underlying probabilistic reasoning behind the computations. It is helpful to understand this to better grasp how and why these algorithms work.

The probability calculation that is at the heart of these algorithms is the likelihood ratio, or odds ratio. We had come across this earlier when discussing DNA profiles and their use in courtroom cases as evidence against a suspect.

In general, a likelihood ratio takes the following form:

$$LR = \frac{\text{P(Observed Data | Alternative hypothesis)}}{\text{P(Observed Data | Null hypothesis)}}$$

If this ratio is large, it is a sign that the alternative hypothesis is more likely, and if it is small, it is a sign that the null hypothesis is more likely.

In our sequence alignment problem, the null hypothesis is that the two sequences are two random, unrelated sequences. The alternative is that the two sequences are related in some way. What we realize is that even if two sequences are unrelated, they could potentially match up well in certain regions just by pure chance. We want to be sure that the amount of matching and alignment we found is due to more than just pure chance.

Also, in the sequence alignment problem, the "observed data" referred to in the likelihood ratio calculation are the two sequences and their alignment. In the discussion that follows, we will simplify things and assume that the two sequences are of the same length and we will not introduce any gaps in the alignment.

Now, let's be more specific with the numerator and denominator terms in the LR with regard to the sequence alignment problem. First, we need some notation. Let our two sequences be represented by **S** and **T**, where each is a sequence of letters (either {A,C,G,T} if they are DNA sequences, or the twenty amino acid codes if they are proteins). We'll refer to the individual residues in **S** as $(s_1, \ldots, s_n)$ and the residues in **T** as $(t_1, \ldots, t_n)$, where n is the length of the sequences. So for example, if **S** is the DNA sequence AATGC, then s_1 =A, s_2 =A, s_3 =T, s_4 =G, and s_5 =C, with $n = 5$.

First, let's look at the denominator of the likelihood ratio. Under the hypothesis that the two sequences are just two random, unrelated sequences, we can calculate the denominator probability as follows (we'll leave off the piece after the conditional symbol for brevity):

$$\text{P(Observed sequences | Random,unrelated)} = \text{P}\begin{pmatrix} s_1 & s_2 & \cdots & s_n \\ t_1 & t_2 & \cdots & t_n \end{pmatrix}$$

The symbols inside of the probability statement on the right represent the alignment of residues across the two sequences (i.e., s_1 is aligned with t_1 and so on).

We will also assume that the bases found in different positions are independent of one another, which is not too far from true for most DNA and protein sequences. With this assumption, we can further write the above probability as

$$\text{P}\begin{pmatrix} s_1 & s_2 & \cdots & s_n \\ t_1 & t_2 & \cdots & t_n \end{pmatrix} = \text{P}\begin{pmatrix} s_1 \\ t_1 \end{pmatrix}\text{P}\begin{pmatrix} s_2 \\ t_2 \end{pmatrix}\cdots\text{P}\begin{pmatrix} s_n \\ t_n \end{pmatrix}$$

Now, under the hypothesis that these are random and unrelated sequences, the probability that s_i and t_i will be in alignment is just the product of their individual probabilities, P_{s_i} and P_{t_i}. For example, if I use a computer to randomly generate the two DNA sequences CTCA and ATAG, the probability that we have a C in position 1 of the first sequence and an A in position 1 of the second sequences is just $(P_{C1})(P_{A2})$ where P_{C1} is the probability in my computer program that I set for randomly selecting a C in the first sequence, and P_{A2} is the analogous probability for the second sequence.

Applying this to the previous equation now gives us

$$\mathrm{P}\binom{s_1}{t_1}\mathrm{P}\binom{s_2}{t_2}\cdots\mathrm{P}\binom{s_n}{t_n} = P_{s_1}P_{t_1} \times P_{s_2}P_{t_2} \times \cdots \times P_{s_n}P_{t_n} = \prod_{i=1}^{n} P_{s_i}P_{t_i} \tag{5.1}$$

As a quick example, let's say those two sequences above (CTCA and ATAG) were generated by a computer program using the following rules: bases A,C,G,T were chosen for the first sequence with probabilities .2,.2,.3,.3 respectively, and were chosen for the second sequence with probabilities .25,.25,.25,.25 respectively. Then, using our above calculations:

$$\begin{aligned}\mathrm{P}(\text{Observed sequences} \mid \text{Random,unrelated}) &= (.2)(.25) \times (.3)(.25) \times (.2)(.25) \times (.2)(.25) \\ &= 0.000009375\end{aligned}$$

Now, let's look at the numerator of the likelihood ratio. Recall that in that case, the underlying hypothesis is that the sequences are related. So, we'd expect the probability of various pairs of bases or amino acids lining up to be more or less than the probability calculated based on random chance. For example, if two protein sequences have a fairly recent common evolutionary origin, and the original sequence was KLVYC, then we'd expect there to be a fairly large probability that the two evolved sequences would both have K's in the first position, both L's in the second position, and so on. This would certainly be more likely than the two sequences just being random selections of the twenty amino acids.

So, to calculate the LR in the numerator, the main difference is that we can't make the simplification that we made in (5.1) for the denominator probability. But otherwise, the development is the same, including the assumptions of independence between residues. The calculation for the numerator is

$$\begin{aligned}\mathrm{P}(\text{Observed sequences} \mid \text{Related}) &= \mathrm{P}\binom{s_1}{t_1}\mathrm{P}\binom{s_2}{t_2}\cdots\mathrm{P}\binom{s_n}{t_n} \\ &= P_{s_1t_1}\ P_{s_2t_2}\cdots\ P_{s_nt_n} \\ &= \prod_{i=1}^{n} P_{s_it_i}\end{aligned}$$

Now we can put together the pieces of the LR. We have

$$\begin{aligned}LR &= \frac{\prod_{i=1}^{n} P_{s_it_i}}{\prod_{i=1}^{n} P_{s_i}P_{t_i}} \\ &= \prod_{i=1}^{n}\left(\frac{P_{s_it_i}}{P_{s_i}P_{t_i}}\right)\end{aligned}$$

It is common to take the log of the likelihood ratio in order to make calculations. Notice that this doesn't change any implications (since the log function is always increasing), but can make computations easier even on a computer because taking sums of probabilities is more efficient than multiplying many together. Taking the log, we get

$$\log LR = \sum_{i=1}^{n} \log\Big(\frac{P_{s_i t_i}}{P_{s_i} P_{t_i}}\Big)$$

We call this the log-likelihood ratio, or the log-odds. Notice that it is a sum over all positions in the sequences of the log of the ratio of two probabilities. This is what the scoring algorithms are based on. Typically, the base used for the log is 2, as this gives an answer that can be interpreted in bits.

It is important to realize that in our alignment algorithms, when we assign a score to a match or a mismatch (such as using a score of 2 for a match between DNA sequences), what is really happening is that we have decided that the log-odds of that match is 2. And scoring a mismatch as -1 says that we are stating the log-odds for that mismatch is -1. In other words, any scoring system is just a way to assign values to the possible log-odds values that get added together in the above log-likelihood ratio.

As a more concrete example, let's take a closer look at the BLOSUM50 matrix. First, we must state that the values in the BLOSUM50 matrix are not exactly the log-odds values, but have been scaled by the factor 3 and rounded off to make them easier to deal with. It should be noted that the scale factor of 3 is not standard, and other scale factors are used for other scoring matrices (e.g., BLOSUM62 uses 2 as the scale factor).

Notice the value assigned when amino acid R is aligned with itself is +7 in this matrix. This means, in an alignment algorithm, if we align an R with an R, we add 7 to the score for that alignment. What it is more directly saying from a probability standpoint is

$$3\log_2\Big(\frac{P_{RR}}{P_R P_R}\Big) = 7$$

(recall that the 3 is just a scale factor that is used). From this, we can directly calculate the value for P_{RR} that the matrix infers, if we make an assumption about the value of P_R. Notice that the fact that a positive number is used here (i.e., 7), suggests that we believe that R will align with itself more often in related sequences than in unrelated sequences. Compare this to the BLOSUM50 matrix value of -4 for I aligning with D. This suggests that such an alignment is much less likely in related sequences than unrelated sequences (as we'd expect).

Thought provoking exercise: Although we haven't discussed this in detail, the scoring matrix that would be used for a particular alignment problem depends on how related the sequences are believed to be. The BLOSUM50 matrix would be used, for example, for proteins that are thought to have a more distant evolutionary origin (more time has passed since the sequences diverged from each other), while the BLOSUM62 matrix would be used for proteins that are thought to have a more recent evolutionary origin (less time has passed since the sequences diverged from each other). With this in mind, would we expect the values in the BLOSUM50 matrix or the BLOSUM62 matrix to be higher in absolute value (regardless of the scale factor used)?

If you are interested, a good readable reference describing the development of the BLOSUM matrices and the PAM matrices is in the Ewens and Grant reference (sections 6.5.2 and 6.5.3).

5.4 Markov Chains

Markov chains are a powerful mathematical tool that allow one to build probability models for very complex processes. We will only cover them briefly in this course, but you should realize that there are full courses, many books, and

a well-developed theory for Markov Chain models. They are particularly useful for many biology-related problems, so it is useful to have at least a starting knowledge of them through this course. A more complete understanding would require knowledge of other areas of mathematics such as linear algebra and more complex probability concepts (which are not required for this course).

5.4.1 Basics

The general idea of a Markov chain model is to describe a process that moves from *state* to *state* in discrete steps (we will only discuss discrete Markov chains). A non-biological example of such a process might be a particular stock you are watching on the stock market. The process in that case is the working of the stock market itself, and the discrete steps we might consider are days. To keep things simple, we might describe the possible states the process can be in as either "stock price went down today", "stock price went up today", or "stock price stayed the same today".

Let's refer to these three states as D, U, and S (for down, up, and same). Then, as we watch the process unfold over the course of, say, two weeks (10 trading days), we can write down the sequence of states the process visited. For example, we might notice the following sequence: DDDUSUUUDD. This tells us that the stock went down in price the first three days, then up, then the same, then up for three days, then finished by going down for two days.

Notice that the order of the sequence is extremely important. The sequence USUDUDDDUD describes a completely different outcome despite the fact that the stock went up, down, and same the same total number of days.

Also notice that the observed sequence was just one of many possibilities. Before the beginning of that 10 day period, we might not have much of a clue about what sequence was about to unfold. If we were about to invest in that stock, we might hope for UUUUUUUUUU, but we would not know for sure if that would happen.

Of course, a next logical question then is to ask, "what is the probability of the sequence being UUUUUUUUUU?" Markov chains give us a framework for building a probability model for such processes and calculate such probabilities (as well as many more interesting and complex questions about the process).

5.4.2 Notation and Setup

First, let's use the symbol X to denote the process, or in other words, the sequence of states in the chain. The individual states visited by the chain will be denoted by $X = (x_1, x_2, x_3, \ldots, x_n)$. Each of the x_i are symbols that represent the possible states in the chain (such as U, D or S in the stock market example). We can write the probability of observing this sequence using standard probability notation and the conditional probability law:

$$\begin{aligned} P(X) &= P(x_1, x_2, \ldots, x_n) \\ &= P(x_n|x_{n-1}, x_{n-2}, \ldots, x_1)\ P(x_{n-1}|x_{n-2}, \ldots, x_1)\ P(x_{n-2}|x_{n-3}, \ldots, x_1) \cdots\ P(x_2|x_1)\ P(x_1) \end{aligned}$$

This just describes the probability of observing a particular sequence in the following way: starting from the right side, the overall probability is the chance of observing the state x_1 first, times the chance of the process being in state x_2 second, given that it was first in x_1, times the chance of the process being in state x_3 next, given that it first visited x_1 and x_2, and so on.

As it stands, this is actually a difficult thing to calculate because there are many complex conditional probabilities involved. A *first-order* Markov chain simplifies things by assuming that the chance that the process will be in a certain state at a certain point has only to do with the state it was in previously. Only the current state has anything to do with what state it will visit next. With this in mind, the above probability reduces to

$$\begin{aligned} P(X) &= P(x_1, x_2, \ldots, x_n) \\ &= P(x_n|x_{n-1})\ P(x_{n-1}|x_{n-2})\ P(x_{n-2}|x_{n-3}) \cdots\ P(x_2|x_1)\ P(x_1) \end{aligned}$$

Now, the calculation is much simpler. All of the probabilities involved (other than the first) can be described as "the probability of the process visiting state k next given that it is currently in state j."

So, in order to completely define this process, we need to write down these probabilities. They are referred to as *transition probabilities* because they represent the probability that the process will transition from one state to another at the next step. We collect all of these probabilities in a matrix which we call the *transition matrix* P for the process. This matrix is a square matrix in which each row represents a possible state of the chain, as does each column.

Let's notate the possible states by $S_1, \ldots, S_k$. The entry in the (i, j)th cell of the matrix, which we'll notate by p_{ij}, represents the probability that the process will transition from state S_i to state S_j at any particular time step. In general, the matrix takes the following form

$$P = \begin{array}{c} \\ S_1 \\ S_2 \\ \vdots \\ S_k \end{array} \begin{array}{c} \begin{array}{cccc} S_1 & S_2 & \cdots & S_k \end{array} \\ \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ & \vdots & & \\ p_{k1} & p_{k2} & \cdots & p_{kk} \end{pmatrix} \end{array}$$

Notice that the probabilities must add to one across each row since the process must transition somewhere at the next step.

We also need one other piece of information: the initial probabilities. The transition matrix itself doesn't tell us the probabilities associated with the initial state the chain will visit (since there is no previous state to base this information on). These initial probabilities will be $p_{01}, \ldots, p_{0k}$, and these must also add to 1.

As an example, let's think again about the stock market process discussed above. Below is a transition matrix we might use to model this process (notice the rows add to 1):

$$P = \begin{array}{c} \\ \text{D} \\ \text{U} \\ \text{S} \end{array} \begin{array}{c} \begin{array}{ccc} \text{D} & \text{U} & \text{S} \end{array} \\ \begin{pmatrix} .5 & .3 & .2 \\ .3 & .6 & .1 \\ .4 & .4 & .2 \end{pmatrix} \end{array}$$

This matrix suggests, for example, that if the stock price was down on a particular day, there is a 50% chance it will be down again the next day, a 30% chance that it will be up the next day, and a 20% chance that it will stay the same the next day. We can make similar interpretations for each row.

For the initial probabilities for this process, we might use 1/3, 1/3, 1/3 to represent the fact that the initial state will be unknown to us.

All of this information gives us everything we need to fully understand and analyze this Markov process. We can, for example, calculate the probability of the sequence DDDUSUUUDD that we listed above. It is

$$\begin{aligned} \text{P(DDDUSUUUDD)} &= \text{P(D)P(D | D)P(D | D) P(U | D) P(S | U) P(U | S) P(U | U) P(U | U) P(D | U) P(D | D)} \\ &= (1/3)(.5)(.5)(.3)(.1)(.4)(.6)(.6)(.3)(.5) \end{aligned}$$

which is a small number, but that is to be expected considering there are $3^{10} = 59,049$ different possible sequences this process could have followed (three possible states at each of ten steps). More interesting would be comparing this probability to the probability of other possible sequences.

Finally, it is possible to incorporate the initial probability distribution right into the transition matrix. We can do this by adding a new state to the process that we'll call the *begin state* and will label as **B**. The process will always start in this state and will leave it immediately without returning. Our transition matrix for the stock market example could then be viewed as

$$P = \begin{array}{c} \\ \mathbf{B} \\ \text{D} \\ \text{U} \\ \text{S} \end{array} \begin{array}{c} \begin{array}{cccc} \mathbf{B} & \text{D} & \text{U} & \text{S} \end{array} \\ \begin{pmatrix} 0 & 1/3 & 1/3 & 1/3 \\ 0 & .5 & .3 & .2 \\ 0 & .3 & .6 & .1 \\ 0 & .4 & .4 & .2 \end{pmatrix} \end{array}$$

5.4.3 Modeling DNA Sequences as Markov Chains

With this quick background on Markov chains, we can now turn back to bioinformatics and take a look at a couple of simple examples of their use in the field. Clearly, we can view a DNA sequence (or an amino acid sequence) as a Markov chain. The process is the DNA sequence itself, and the steps are represented by the bases as we move in a linear fashion from the 5' to 3' end of the DNA. The states, of course, are the four bases A, C, G, and T. In general, our transition matrix would look like the following

$$P = \begin{array}{c} \\ \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \end{array} \\ \begin{pmatrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{pmatrix} \end{array}$$

where p_{ij} is the probability that base i will be followed by base j. The rows, as always, sum to 1.

Typically, what we will do is estimate these probabilities from real DNA sequences to help us better understand the process. For example, take the following DNA sequence that we saw from Assignment 4 (part of the HBB gene in humans)

```
1 acatttgctt ctgacacaac tgtgttcact agcaacctca aacagacacc atggtgcatc
```

```
 61 tgactcctga ggagaagtct gccgttactg ccctgtgggg caaggtgaac gtggatgaag
121 ttggtggtga ggccctgggc aggctgctgg tggtctaccc ttggacccag aggttctttg
181 agtcctttgg ggatctgtcc actcctgatg ctgttatggg caaccctaag gtgaaggctc
241 atggcaagaa agtgctcggt gcctttagtg atggcctggc tcacctggac aacctcaagg
301 gcacctttgc cacactgagt gagctgcact gtgacaagct gcacgtggat cctgagaact
361 tcaggctcct gggcaacgtg ctggtctgtg tgctggccca tcactttggc aaagaattca
421 ccccaccagt gcaggctgcc tatcagaaag tggtggctgg tgtggctaat gccctggccc
481 acaagtatca ctaagctcgc tttcttgctg tccaatttct attaaaggtt cctttgttcc
541 ctaagtccaa ctactaaact gggggatatt atgaagggcc ttgagcatct ggattctgcc
601 taataaaaaa catttatttt cattgc
```

We can count up the number of times each bases follows each other base and convert to percentages to get the following transition matrix

$$P = \begin{array}{c} \\ \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \end{array} \\ \begin{pmatrix} .2993 & .2628 & .2482 & .1898 \\ .2821 & .2756 & .0385 & .4038 \\ .1879 & .2667 & .3152 & .2303 \\ .1198 & .2036 & .4371 & .2395 \end{pmatrix} \end{array}$$

We notice, for example, that a CG pair is somewhat uncommon (as we have seen in an earlier example in the course, and will discuss further below), while a T followed by an G is more common than might be expected. These insights could be an effect of this particular DNA sequence and the DNA region it comes from, or it could be a more global effect true of DNA sequences in general.

5.4.4 CpG Islands

This example is taken from Durbin, et al. It relates to the fact, which we have now seen a couple of times, that the bases CG tend to not appear together in that order nearly as much as we'd expect. Here is an excerpt from Durbin regarding the situation and a question we can investigate:

> In the human genome wherever the dinucleotide CG occurs (frequently written as CpG...), the C nucleotide is typically chemically modified by methylation. There is a relatively high chance of this methyl-C mutating into a T, with the consequence that in general CpG dinucleotides are rarer in the genome than would be expected from the independent probabilities of C and G. For biologically important reasons the methylation process is suppressed in short stretches of the genome, such as around the promoters or 'start' regions of many genes. In these regions we see many more CpG dinucleotides than elsewhere, and in fact more C and G nucleotides in general. Such regions are called CpG islands (Bird, 1987). They are typically a few hundred to a few thousand bases long.

The question posed at the end is: if we have a DNA sequence for which we don't know whether or not it is a CpG island or not, what techniques can we use to decide if it is? The answer is based on analysis of two Markov chain processes and a likelihood ratio test, as we'll see below.

First, in order to do this, we need to have data from real DNA sequences which we can classify as either coming from a CpG island or not. For those sequences which we know to come from a CpG island, we'll estimate the transition probabilities of a Markov chain. Separately, for those sequences which we know to not come from a CpG island,

we'll estimate the transition probabilities of a second Markov chain. These two chains are modelling two different processes: one models DNA sequences from CpG islands, and the other models DNA sequences from non-CpG islands.

Data from 48 sequences were used by Durbin, et al., and each of these sequences was known as either a CpG island or not. The resulting transition matrix estimates were

CpG Islands

$$P = \begin{array}{c} \\ \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \end{array} \\ \begin{pmatrix} .180 & .274 & .426 & .120 \\ .171 & .368 & .274 & .188 \\ .161 & .339 & .375 & .125 \\ .079 & .355 & .384 & .182 \end{pmatrix} \end{array}$$

Non-CpG Islands

$$Q = \begin{array}{c} \\ \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \end{array} \\ \begin{pmatrix} .300 & .205 & .285 & .210 \\ .322 & .298 & .078 & .302 \\ .248 & .246 & .298 & .208 \\ .177 & .239 & .292 & .292 \end{pmatrix} \end{array}$$

A quick look at these matrices shows some clear differences between these two types of regions in DNA sequences, particulary with regard to the CG sequence.

Now, suppose we have a new DNA sequence $\mathbf{X} = (x_1, \ldots, x_n)$ which is an observed sequence from a Markov chain, and we'd like to know whether or not it seems to be a CpG island or not. This would help us determine whether the sequence is a promoter region, or some other area of DNA. For example, let's again use the 626 base pair sequence above from the HBB gene.

Our method for deciding will again be the idea of a log-likelihood ratio. We have two hypotheses here: either this sequence comes from a CpG island or not. We can calculate the likelihood that this sequence comes from a CpG island and the likelihood that this sequence does not, and take the ratio. Since we have two well-defined models above, one for each hypothesis, we can calculate these values. More specifically, our calculation will be:

$$\log LR = \log \frac{\text{P}(\mathbf{X} \mid \text{CpG Island})}{\text{P}(\mathbf{X} \mid \text{non-CpG Island})}$$

$$= \log \frac{p_{x_0 x_1} p_{x_1 x_2} p_{x_2 x_3} \cdots p_{x_{n-1} x_n}}{q_{x_0 x_1} q_{x_1 x_2} q_{x_2 x_3} \cdots q_{x_{n-1} x_n}}$$

$$= \sum_{i=1}^{n} \log\Big(\frac{p_{x_{i-1} x_i}}{q_{x_{i-1} x_i}}\Big)$$

This says that for each base in this new sequence, we need to calculate the transition probability to that base from the CpG island model and from the non-CpG island model, take the ratio, take the log, then add up over all the bases. As we have seen before, a larger value of this likelihood ratio suggests that the hypothesis in the numerator (namely that this sequence comes from a CpG island) is more correct, and vice versa if it is a small value.

We can make this calculation for the HBB sequence. Recall that the first few bases in this sequence are

```
acattt....
```

So the log-likelihood ratio calculation is (using .25 as the initial probabilities to start the chains)

$$\log LR = \log(\frac{.25}{.25}) + \log(\frac{.274}{.205}) + \log(\frac{.171}{.322}) + \log(\frac{.120}{.210}) + \log(\frac{.182}{.292}) + \log(\frac{.182}{.292}) + \cdots$$

$$= -87.84$$

Since this value comes out negative, it suggests that the non-CpG island model does a better job of modelling this DNA sequence, and therefore suggests that the sequence comes from a non-CpG island. In this particular case, we know this to actually be correct because most of the sequence is in the coding region of the HBB gene.

5.4.5 Testing for Independence of Bases

We will introduce another example here, but leave the details for an assignment. We have discussed earlier in the course the idea of testing whether the bases in a DNA sequence are independent of one another. In a number of situations, we have had to make this independence assumption to make the mathematics easier and complete the problem without making it overly difficult. However, it is not always true, and in certain regions of DNA, it is certainly not true.

Markov chains give us a simple way of testing for independence between bases and modeling longer range dependence among pairs of bases in a DNA sequence. For example, we can build a Markov chain model that allows for the base at a particular position to depend on the bases in the previous position, the previous two positions, or in general, any number of previous positions. If you think carefully about such a model, you will notice that it can still be described by a Markov model, but not necessarily a first-order model. The longer range of dependence we want to account for, the higher order Markov model we will need.

As we have seen, a first-order model uses four states (the four DNA bases) and defines the probability of observing a base at a certain position as dependent only on the base in the previous position. In the homework assignment, you will develop a method for analysing independence based on this framework.

It turns out that we can treat higher order models as a first-order model as well if we redefine the states carefully. For example, suppose we wish to model a situation where the based found at a position depends on the previous *two* bases. This requires a second-order model if we continue to define the state space as consisting of the four DNA bases. However, if we redefine the state space to consist of all possible pairs of bases (i.e., 16 states such as AA, AC, AG, etc.), we can now think of this as once again a first-order model, but with $4^2 = 16$ states. Even though there are more states, this is often the simpler way to think about it because we can now rely on all the first-order Markov

chain model theory.

As an example of redefining the states, consider the following DNA sequence:

```
acggcgtta
```

If we define the 16 states as all possible pairs of bases, we can now view this sequence as the following sequence of states in this new state space:

```
ac-cg-gg-gc-cg-gt-tt-ta
```

The initial state is ac, the chain then transitions to the state cg, and so on. You should note that this definition of states is such that some transitions are impossible. You will investigate this more in the homework assignment.

5.5 Other Sequence Problems in Bioinformatics

There are many other categories and subcategories of problems relating to DNA or protein sequence analysis in bioinformatics. We will discuss a couple others here that are particularly interesting and solvable using techniques we have discussed in the course.

5.5.1 Gene Finding

Identifying genes is an important problem in bioinformatics. In a sentence, the problem can be stated as the following: "Given a long sequence of DNA from a certain genome, determine the portions of that sequence that encode for a gene or genes." If we can do this, we can then translate the coding sequence into an amino acid sequence and further study its function.

In this section, we will discuss some basic concepts in gene finding (also called gene prediction), including the difference between finding genes in prokaryotic versus eukaryotic genomes and the types of signals that can be searched for.

Prokaryotic versus Eukaryotic Organisms

At a very high level, organisms are classified as either *prokaryotic* or *eukaryotic*. Prokaryotic organisms are those whose cells do not have a nucleus. This set of organisms essentially comprise bacteria and are often single-celled. Eukaryotic organisms are those whose cells do have a nucleus. Eukaryotes comprise all other higher forms of life, from algae, fungi, to plants, animals, and of course humans. Although they are simple organisms, prokaryotes (bacteria) are well-studied because of the relative ease of genome analysis (their genomes are small and contain many tightly packed genes), easy to manipulate, and because of their affect on humans.

The main biological difference that results from not having a nucleus is that genes in prokaryotes generally do not contain introns. In other words, when a gene is transcribed into mRNA, it is translated into protein in that same form. As we have seen, eukaryotic mRNA goes through another level of processing while still in the nucleus, namely, intron removal. The spliced mRNA containing only the exons is then transferred out of the nucleus and only then translated into protein.

Gene Finding in Prokaryotes

Because of the difference discussed above, gene finding in prokaryotes is much simpler than in eukaryotes. We will discuss this first.

First, we should note that many bacterial genomes have been completely sequenced (such as the common bacteria *E. coli*. They are relatively small genomes (on the order of 1 to 5 million bases) and have relatively few genes (on the order of a few thousand). In the cases where a complete sequence is available, this can be used as the basis for gene finding. In other cases, there typically is access to partial sequences on which we search for genes.

In prokaryotes, the process boils down to the following: given a stretch of DNA, search for a the start codon AUG and follow it to the first stop codon, one of UAA, UAG, UGA (using the RNA notation of U instead of T). Since no intron removal is done, that sequence is a gene and can directly be translated into the corresponding amino acid sequence using the genetic code.

This sounds very easy. It isn't too difficult, but there is a bit more to consider. Mainly, every AUG triplet found is not a start codon and every UAA, UAG, or UGA is not a stop codon. For example, the AUG sequence you find may actually span two codons, as in GGA-UGC. So, we have to consider that there are different *open reading frames*, or *ORFs*, in which we can read, or make sense of the DNA sequence.

In fact, there are six possible ORFs for any stretch of DNA. Three are from the fact that we can begin reading and translating at either the first, second, or third available base. Each of these frames leads to a different sequence of codons and therefore a different sequence of amino acids. Note that starting at the fourth base is equivalent to starting at the first because codons are triplets of bases.

In fact, there are three other ORFs in a stretch of DNA. To realize this, we need to once again recall that DNA actually comes in double stranded form, with complementary bases lining up across the strands. Biologically, DNA is read, transcribed, and translated in the 5' to 3' direction. This can happen on either strand. The 5' to 3' direction on the complementary strand runs in the opposite direction. So for a particular stretch of double stranded DNA, there could be a gene on one strand reading left (5') to right (3'), or one on the other strand reading right (5') to left (3'). The figure below helps to show this.

```
3'  UAAGUACGCAAUACGGCUUGGGCGUAGGCAAUC 5'
5'  AUUCAUGCGUUAUGCCGAACCCGCAUCCGUUAG 3'
```

So, the second set of three ORFs are read from the 5' to 3' direction on the complementary strand, giving a total of six possible open reading frames in a single stretch of DNA.

When analyzing each ORFs, we will search for a pattern where we see a start codon, a long stretch with no stop codons, and then finally a stop codon. Typically genes are not short - usually at least 300bp. So, this would be a sign that we are reading in the correct frame and have found a true gene.

On the other hand, if we are reading in the incorrect frame, we would likely come across an AUG triplet, then find one of the stop codons not too far downstream (since the appearance of UAA, UAG, or UGA in a random sequence is likely to occur at some point). For this to be a real gene, it would be an unusually short one, and so

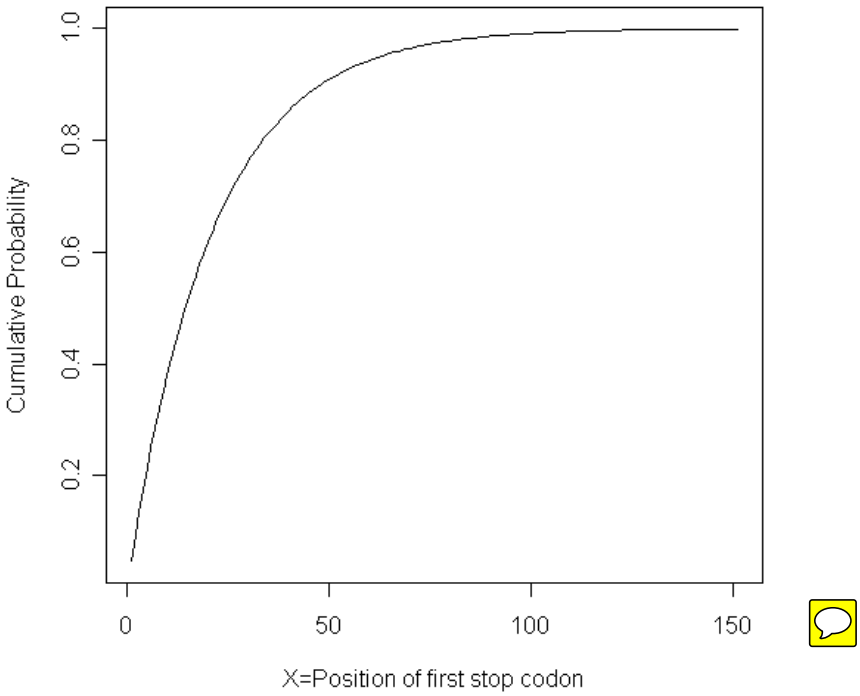
we might conclude that it is not really a gene and just a region of non-coding DNA or is being read in the wrong ORF.

We can make some probabilistic calculations to support the above contention. Suppose we are reading in a particular ORF that is incorrect and not encoding a gene. We can consider the base sequence, and the resulting codon sequence, as being completely random. Consider a point where we come across the AUG start codon. In this completely random sequence, how long would it take to come across a stop codon?

As a simple analysis of this problem, we might assume that each of the 64 codons is equally likely to occur at any position. In that case, since there are three stop codons, we can say that any particular random codon has a 3/64 chance of being either UAA, UAG, or UGA. We can then model the random variable that records the first appearance of one of these three codons as a geometric random variable:

$$X \sim \text{Geometric}(p = 3/64)$$

where X=the codon position where UAA, UAG, or UGA first appears (following an AUG) in a completely random sequence. We can calculate the expected value of X as $64/3 = 21.3$, and a picture of its cumulative distribution function is below.

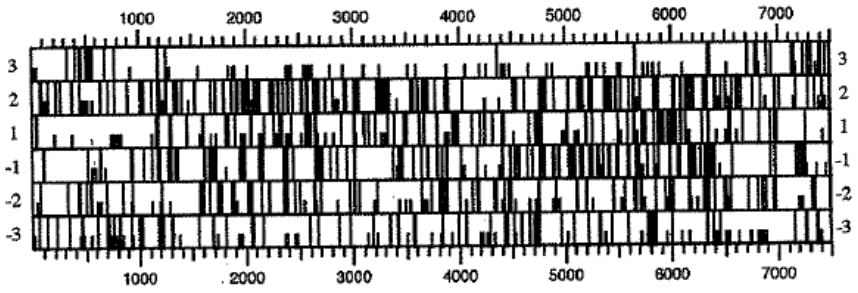


The first such codon is highly likely to occur ($> 90\%$ probability) prior to the 50th codon, and there is approximately a 99.2% probability that it will occur before the 100th codon. Therefore, if we are reading an ORF and find an apparent short gene, it give strong credence to the hypothesis that we are just reading a random DNA sequence and have not found a true gene, since most real genes are known to be over 100 codons long.

However, of course, we could make a mistake. On one hand, there are bacterial genes shorter than 100 codons. So we would incorrectly mislabel them as not being genes according to the above analysis because we would assign a high probability that it arose from random junk DNA. Secondly, there is a small chance that a random sequence of DNA will have an apparent gene of length over 100 codons (about .8% according to our geometric model above). We might mislabel that region as an actual gene according to this analysis. However, if we find a very long "apparent gene", say 200+ codons, we can be almost certain that it is a real gene since the probability that a random DNA sequence will need 200+ codons before encountering a stop codon is so small.

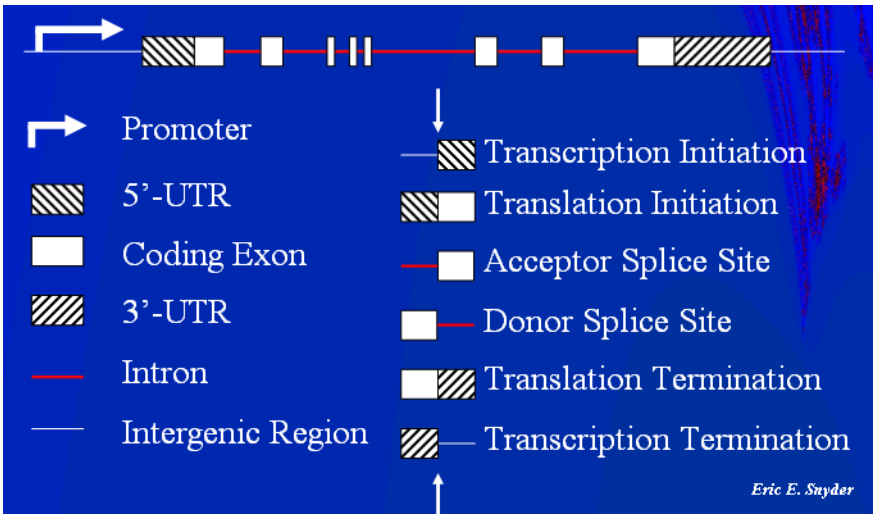
The other caveat to our simple analysis here is that 3/64 is a simple estimate of the probability that a codon will be one of the three. A more complete analysis would take into account other factors such as the underlying probabilities of each base in the particular genome, and any dependencies among bases from one position to the next.

A nice picture of a real gene search in *E. coli* over a stretch of 7500 bp is given below (from Mount, 2004). The six open reading frames are shown, with the position of AUG codons displayed as short bars and the position of stop codons shown as full bars. Notice the wealth of stop codons throughout most of the reading frames except the one labeled "3" at the top. In that frame, there is a start codon at position 1284 and we don't encounter the first stop codon until position 4355. This is a strong gene signal, one having length of nearly 3100 bp. In fact, this is known to be the gene called *lacZ* in *E. coli*.



Gene Finding in Eukaryotes

As we have mentioned, finding genes in eukaryotes (e.g., humans) is more difficult than in prokaryotes because a DNA sequence is not translated directly into an amino acid sequence. Introns are first removed. So, we would have to know where the introns are prior to making this translation. A nice picture from http://obesitygene.pbrc.edu/~eesnyder/presentations/GeneParser_Presentation_files/frame.htm (Snyder and Chepudira) shows the basic structure of eukaryotic gene regions.



Starting with a given eukaryotic DNA sequence, there are two basic steps to determining if there is a real gene located in the sequence and where it is:

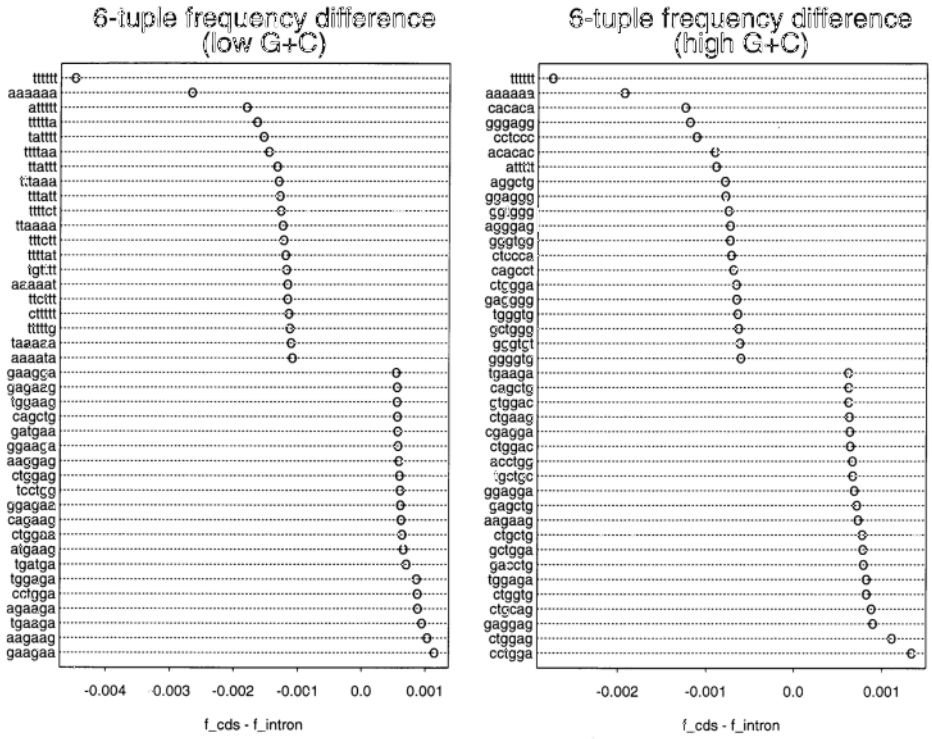
1. Predict where the introns are and remove them from the sequence. Or, equivalently, predict the exons directly to build the sequence.
2. Now treat the sequence as we did with prokaryotes: search different ORF's for characteristic patterns of real genes.

In each case, there are a number of methods and techniques that have been developed, and doing these correctly typically requires much time and many iterations. We will cover a couple of useful methods.

We'll begin with the first item, detecting exons. This problem is particularly hard, especially in higher eukaryotes where a typical gene consists of a number of small exons that are separated by large introns, typically 10 to 100 times larger than the exons (Zhang, 1998). The methods for solving this problem are very complex, so we will focus on some basic ideas that are related to topics we have covered in the course.

Two key facts that we will make use of have been observed about genes and exons in different organisms. One is that the usage of the 61 protein-encoding codons varies from organism to organism, even for the codons that redundantly code for the same amino acid. We saw this before, and referred to the codon usage data at `http://www.kazusa.or.jp/codon` for different species. This data represents frequencies of codons in real genes. We can make use of this to score a potential exon as to whether the encoded codons make sense for that organism.

Similarly, it has been found that contiguous sequences of six bases (called *hexamers* or *6-mers*) have characteristic frequencies within gene coding regions, and these frequencies are different from organism to organism. An example of these differences in humans from Zhang, 1998 is shown below. Not all the possible hexamers are shown in this chart, just the top and bottom 20. The top 20 are those that occur most frequently in exons compared to introns, and vice versa for the bottom 20. Notice that the top 20 are typically hexamers with repeated bases, such as tttttt and aaaaaa. Note that the two different graphs represent differences between genes high in the GC bases and low in the GC bases.



A methodology for using this information is based on GeneParser (Snyder and Stormo, 1995). Let's focus on the codon usage table as an example. We can construct a matrix that has the target DNA sequence along the left side and along the top (similar to the matrix we setup when aligning two sequences). Then, each cell in the matrix, for example the (i, j) cell, represents a log-likelihood ratio score for the likelihood that the DNA fragment from positions i through j comes from an exon compared to an intron. High positive scores are predictive of exon locations, and high negative scores would be predictive of intron locations. Assembling the best predictive sequence of exons and introns is complex, but can be based on these scores.

We can also construct similar matrices based on other statistical measures that contrast known exons to known introns, like hexamer frequencies, dinucleotide frequencies, and length distributions (see Zhang 1998 for more data on these statistics).

As an example, let's again make use of the codon usage table for humans from http://www.kazusa.or.jp/codon as a representation of the distribution of codons in human exons. The current version of that table is repeated here:

```
UUU 17.4(633626)  UCU 15.1(548455)  UAU 12.1(441486)  UGU 10.5(381094)
UUC 20.4(743002)  UCC 17.7(642035)  UAC 15.3(556798)  UGC 12.6(458929)
UUA  7.6(274788)  UCA 12.2(442404)  UAA  1.0( 37072)  UGA  1.6( 57452)
UUG 12.8(467035)  UCG  4.5(161781)  UAG  0.8( 28728)  UGG 13.2(480244)

CUU 13.1(475663)  CCU 17.4(634220)  CAU 10.8(392988)  CGU  4.6(166773)
CUC 19.7(715343)  CCC 19.9(723912)  CAC 15.1(548066)  CGC 10.6(384290)
CUA  7.1(259841)  CCA 16.9(612909)  CAA 12.2(442229)  CGA  6.2(224492)
CUG 39.9(1449753)  CCG  7.0(254350)  CAG 34.2(1243194)  CGG 11.5(417971)
```

```
AUU 15.8(575464)  ACU 13.0(473799)  AAU 16.8(610977)  AGU 12.1(441137)
AUC 20.9(760429)  ACC 19.0(689901)  AAC 19.1(693831)  AGC 19.4(706723)
AUA  7.4(270016)  ACA 15.0(544170)  AAA 24.2(879684)  AGA 12.0(434655)
AUG 22.1(801969)  ACG  6.1(220632)  AAG 32.0(1163126)  AGG 11.9(432954)

GUU 11.0(399567)  GCU 18.5(672416)  GAU 21.7(789799)  GGU 10.8(392298)
GUC 14.5(528840)  GCC 28.0(1018345)  GAC 25.2(914677)  GGC 22.4(814464)
GUA  7.1(257442)  GCA 15.9(579156)  GAA 28.7(1043166)  GGA 16.5(597986)
GUG 28.3(1028789)  GCG  7.5(271820)  GAG 39.6(1441162)  GGG 16.5(599428)
```

We will use a random equal-codong usage model of codons to represent the distribution of codons in introns (although we could base the intron model on the distribution from known introns). Consider the DNA sequence

```
accgttg
```

The score we would assign in position (1,6) of the matrix would be the log-likelihood ratio corresponding to acc-gtt being an exon as opposed to an intron. The numerator based on the codon usage table would be (assuming independence between codons) $.019 \times .011 = .000209$ and the denominator would be $(1/64)^2 = .000244$ giving a log-likelihood ratio of -0.1554. This is indicative of this region being more likely an intron than an exon.

We could also use Markov chains to make these calculations. Two Markov chain transition matrices can be built to make these calculations, one for the exon model and one for the intron model. These Markov chains could also take into account dependencies from one codon to the next, therefore incorporating the hexamer frequencies directly into the model. For example, looking at the hexamer figure above, the transition probability from gaa to gaa (the bottom line on the left) would be quite small in the exon model to represent the fact that very few exons contain these two codons in succession.

Alternative methods for exon and intron detection include Hidden Markov Models, which are interesting extensions of Markov chains, and neural networks which are a powerful machine learning modeling method.

Now that we have identified exons, we can piece the puzzles of the resulting mRNA together. We still have to search ORFs for start and stop codons and see if our result makes sense. Typically, in eukaryotes, the first AUG in the sequence is the start codon, but other AUGs close to the 5' end but not necessarily the first could be used. We can follow the appropriate ORF through to the first stop codon.

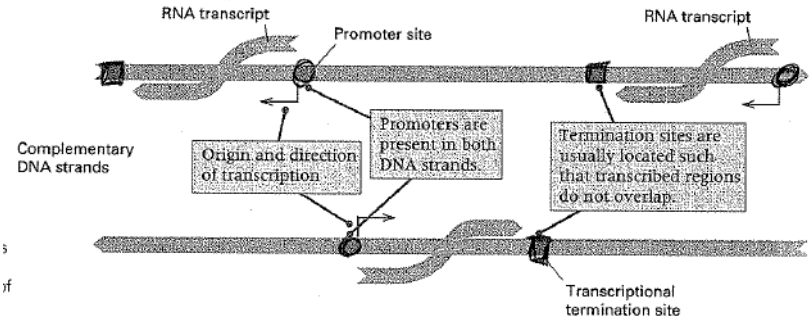
Of course, none of this guarantees that we have found a true gene. The exon detection process is still just a prediction, and is not likely to be perfect. For a real gene-finding problem, the researcher will likely take the predicted gene region (or related protein sequence) and search for homologies using BLAST or similar database similarity search programs. If no homology is found, or only low scoring ones, then we probably made a mistake with our exon detection and piecing together of the mRNA. If we find a strong match, we are probably on the right track.

5.5.2 Identification of Promoter Regions

Background

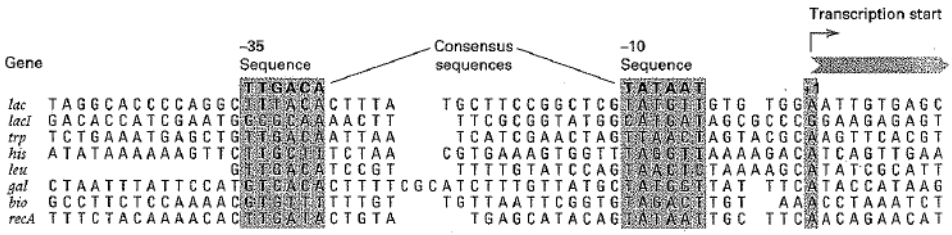
We have briefly mentioned the concept of the *promoter region* of a gene. It is a region upstream of the transcription start site that the RNA polymerase attaches to in order to begin the transcription process. Figure 5.2 gives an overview of the transcription process and related promoter sites.

Figure 5.2: Typical arrangement of promoters, transcripts, and termination sites (Hartl and Jones, 2005)



Typically, the promoter region is defined well by a certain DNA sequence pattern, or combinations of DNA patterns. For example, in many *E.coli* genes, there is a sequence about 35 bases upstream of the transcription start site, called the *-35 sequence*. Not all genes have the same sequence pattern at that point, but if you analyze the -35 sequence across many *E.coli* genes, the *consensus sequence* found is TTGACA. The consensus sequence is determined by taking the base that occurs the most at each position when analyzed over many genes. Also in *E.coli*, there is a region 10 bases upstream of the transcription start site called the *-10 sequence* or the *TATA box* which also serves as a promoter region. The consensus sequence is TATAAT. Figure 5.3 gives an example of these regions for eight genes in *E.coli*.

Figure 5.3: Eight *E.coli* genes showing variable promoter sequences at -35 and -10 bp upstream of start site (Hartl and Jones, 2005)



Notice that for any particular gene, the promoter sequence may not be the same as the consensus promoter sequence, and it could be quite different. Close similarity to the consensus promoter sequence has been associated with the strength of binding of RNA polymerase to the promoter region. We would call such promoters *strong promoters* as opposed to *weak promoters*. Strong promoters have a better ability to begin the transcription process and therefore the related gene will get transcribed (and eventually translated into protein) in higher quantities. So, promoter

regions also have an effect on the level of gene expression.

In humans, the TATA box is often found about 20 bases upstream of the transcription start site. But promoter regions, and related DNA sequence patterns, can be very complex. An example of a gene and related promoter region in humans is shown in Figure 5.4. The output is from GenBank associated with sequence accession number X84664, and is the *stromelysin-3* gene. Notice the position of the TATA signal (as it is called in the listing) and the first exon and gene start point in this sequence.

Promoter Finding

As we have discussed when talking about finding genes, it is important to analyze a long DNA sequence, or even an entire chromosome or genome, to find genes. In addition to other methods we have discussed, we can find genes by searching for known promoter regions. If promoter regions sequences were always exactly the same (i.e., perfectly matched the consensus sequence), this would be a simple computer algorithm to search for known consensus sequences. Unfortunately, as we have seen, true promoter sequences can vary from the consensus sequence, and so we have to allow for this in our search.

Let's set the problem up mathematically. We have access to a long DNA search sequence of length L. We'll refer to this sequence as $S = (s_1, s_2, \ldots, s_L)$. This sequence can be a stretch of DNA known to be upstream of a gene or in an intergenic region, or is potentially an entire chromosome or genome. We will scan through this search sequence to search for possible promoter sites. We can then score each position in this sequence as to how likely it is to be the start of a promoter region. High scoring positions will require further investigation.

The first method we will discuss is fairly simple statistically. We will define a promoter target sequence based on the consensus sequence. Let $T = (t_1, t_2, \ldots, t_l)$ represent this consensus sequence with length l. In Figure 5.3, the -35 promoter consensus sequence has $l = 6$ with $T = (TTGACA)$. Notice that this disregards information about the variation between actual promoter sequences that went into defining the consensus, but we'll take this simple approach for now.

Now, when we scan the long search sequence, we will score each position according to how many bases exactly match the target T. We'll call this a *match score*. This generates a score for each of the L positions in the search sequence (except for the last $L - l + 1$ positions for which we do not have enough bases in the target to match). For example, suppose we use the sequence T above from the -35 consensus sequence in *E.coli*, and our search sequence S is as given below (taken from a portion of the *lac* gene in Figure 5.3.

```
TAGGCACCCCAGGCTTTACACTTTA
```

This sequence S is of length $L = 25$. Since our target is length $l = 6$, we can score the first 20 positions of S. Those scores are below, annotated underneath of the search sequence.

```
S:     TAGGCACCCCAGGCTTTACACTTTA
Score: 31211201021011523001-----
```

Notice that the actual promoter site scores a 5 and clearly stands out within this short sequence. However, in general, if we scan a long sequence of length L, we may expect to find high matching to the consensus sequence even within

random non-promoter regions of DNA. So, every "high matching" score does not infer that it must be a promoter region. Also, some low matching scores might be promoter regions. For example, some of the genes in Figure 5.3 have a match score as defined above as low as 2.

We want to be able to attach a significance level to these match scores so that we can focus our attention on the most significant scores. Doing so is fairly straightforward if we make assumptions about what a random DNA sequence looks like. In a simple model of random DNA, we can assume all bases are independent and equally frequent. For a random DNA sequence $S = (s_1, \ldots, s_l)$ of length l under this model, the number of matches we expect to the target sequence follows a binomial distribution with parameters $n = l$ and $p = 1/4$. For $l = 6$, the match score distribution is

Match Score	Probability
0	.178
1	.356
2	.297
3	.132
4	.033
5	.004
6	.000244

So in the *lac* gene example, the p-value associated with finding a match score of 5 is .004+.0002 =.0042. This is small, and we would assume that we have found an interesting site, unlikely to be simply from random DNA.

One difficulty with this method is that if our search sequence is quite long (L is large), then we are likely to find many false positives. In other words, for $L = 1000000$ bp, even in completely random DNA we would expect to find 244 positions with a match score of 6 (perfect match to the target promoter). The longer the target sequence is, the less of a problem this would become. Also, the equal frequencies model is not really appropriate, and so the binomial distribution is not the best choice. We could use a Markov chain to model more complex base probabilities and even dependencies between bases (think about this on your own).

The other difficulty with this simple method is that we have only based the score on the consensus sequence itself, and not its constituent component pieces, namely each individual sequence which produced the consensus. The information on the variation between different promoter regions is often stored in a *position weight matrix*, or *PWM* (Hertz and Stormo, 1999). These are also often called *position specific score matrices*, or *PSSM*.

There are various methods for deriving a PWM. We will discuss a simple one here. Let's continue to base our example on the *E.coli* genes listed in Figure 5.3. First, we can develop an alignment matrix, which simply counts the number of times each base appears at each position among the genes we have information on (in this case we have 8 such genes). For this example, this matrix would be:

	1	2	3	4	5	6
A	0	0	0	5	1	6
C	0	1	1	2	4	0
G	3	0	6	0	0	0
T	5	7	1	1	3	2

Each column adds to 8 since there were 8 sequences.

Now we can derive an associated probability matrix that simple gives the percentage of times each base appeared in each position. This matrix is simple the values in the previous matrix divided by 8 for this example. The columns of this new matrix will add to one. This is the PWM (or at least one version of it - see Hertz and Stormo, 1999).

	1	2	3	4	5	6
A	0	0	0	.625	.125	.750
C	0	.125	.125	.250	.500	0
G	.375	0	.750	0	0	0
T	.625	.875	.125	.125	.375	.250

Now we can score a search sequence at each position according to how well it matches the promoter sequence by using the PWM as our basis for scoring, and the concept of likelihood ratios. For a search sequence $(s_1, \ldots, s_l)$, calculate the log-likelihood ratio as:

$$\log LR = \sum_{i=1,\ldots,l} \frac{P(s_i|\text{real promoter})}{P(s_i|\text{random DNA})}$$

Then high scoring positions according to this calculation are analyzed for further investigation. Determining the significance of such scores is a little more difficult, and often simulation methods are used.

Figure 5.4: GenBank results for stromelysin-3 gene (some parts deleted for brevity)

```
LOCUS       HSSTROM3                1660 bp    DNA     linear   PRI 18-APR-2005
DEFINITION  H.sapiens stromelysin-3 gene.
ACCESSION   X84664
FEATURES             Location/Qualifiers
     source          1..1660
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /map="22q11.2"
                     /clone="111E10"
                     /clone_lib="LL22NC01"
     TATA_signal     1439..1443
     GC_signal       1445..1450
     exon            1470..1599
                     /number=1
     gene            1492..>1599
                     /gene="stromelysin-3"
     CDS             1492..>1599
                     /gene="stromelysin-3"
                     /codon_start=1
                     /product="stromelysin-3"
                     /translation="MAPAAWLRSAAARALLPPMLLLLLQPPPLLARALPP"
ORIGIN
        1 ...
      601 aagctgaaga actggccagt ccctgccata tgcctcactt tcccctggga cacattttaa
      661 tatccctttc ctggccaggt gcaatggctt ccccatgtaa tcccagcact ttgggaggcc
      721 aaggtgggca gatcacttga ggtcaggagt tcgagaccag cctggccaac atggtgaaac
      781 cccatctcta ctaaaaatac aaaaattagc caggcatggt ggcgcacgtc tgtggtccca
      841 gctacttggg aggctgaggt agcagaatcg tttgaacctg ggaggcggag gttgcagtga
      901 gtggagatca caccactgca ctcctgcctg ggagacagag tgagactgtg tctcaaaaaa
      961 taataataat aaaaaataaa ataatatccc tttcctcaca ggggctattg tgtcatcttc
     1021 tagaaggatc cgttgaggct ctgaggggtg ggggaacttg cttgtgggta ggaccacctg
     1081 tcagaggtca gaggtcaggc caccaaggag acccagtggg atgcgccttc caaaggtggg
     1141 ggtacggatg ggacccatga aacctgactc ctctcagact ctagccaagt ctaagacttt
     1201 ggacggccac cacccagagg agaaactgag acccagagcg gcacgggttg gccagggtca
     1261 accagcacca gatagggact ttggcagccc cggggcagga ccctgtctcc ggccctcgac
     1321 cccgctgggc cgtaccctcc ccgttcacct cccccacccg ggccgcggct gctaggagag
     1381 ttcagaacaa aaggcggcgg ggggcggggc cgaggcgggc cgggggtggg gcggaagcta
     1441 taaggggcgg cggcccggag cggcccagca agcccagcag ccccggggcg gatggctccg
     1501 gccgcctggc tccgcagcgc ggccgcgcgc gccctcctgc ccccgatgct gctgctgctg
     1561 ctccagccgc cgccgctgct ggcccgggct ctgccgccgg tgagtgcccg ccactcgccg
     1621 gccgctcctc gctgaggggg cgccgggcac gcgggctggg
```

Chapter 6

Microarray Analysis

6.1 Overview

This chapter focuses on the analysis of *DNA microarray experiments*, or just *microarrays*. The overall goal of such experiments is typically to conduct an analysis of *gene expression* levels across different experimental conditions. This can help us better determine the function of genes, as well as pinpoint genes that are associated with various conditions, such as having breast cancer, having leukemia, or being exposed to toxins. It will also allow us to better understand how genes work together, and in what situations they tend to be "turned on" or "turned off" at the same time.

6.2 Background on Gene Regulation and Expression

First, we need to review some more biological background in order to fully understand the purpose of microarray experiments and how they work. We will discuss the concepts of gene regulation and expression. Our biological discussions up until now have focused on the biological and molecular mechanisms of how genes are transcribed and translated. Now, we want to answer questions like *why*, *when*, and *how often* are they transcribed and translated.

An important concept to realize is that all genes are not active at all times in all cells. In other words, a human cell may have 25,000 genes encoded in the DNA of its constituent chromosomes, but that doesn't mean that all 25,000 take turns being transcribed and translated in that cell. Genes and their resulting products have particular functions, and so their transcription and translation are regulated as necessary to meet the needs of the cell and the organism as a whole. The amount of mRNA or protein present in a cell that result from transcribing and translating a gene are referred to as the *expression level* or *gene expression* for that gene in that cell.

Some genes are called *housekeeping genes*. These genes (more specifically, the resulting RNA and protein products) are necessary to run basic metabolic and cellular processes, and are typically active in all cells. For example, genes that produce RNA polymerases and those that produce transcription factors, both of which play a role in the process of gene transcription, are needed in every cell. Therefore, housekeeping genes are not differentially regulated from cell to cell, and tend to be transcribed/translated at nearly the same rate at all times, in all cells. Attempts have been made to identify all genes that would be considered housekeeping genes (see Eisenberg and Levanon, 2003 and

related website http://www.cgen.com/supp_info/Housekeeping_genes.html).

Some genes are regulated and expressed only in certain cells. For example, brain cells, muscle cells, skin cells, and liver cells all have very different functions that are specific to that particular kind of cell. For example, the genes necessary to make a brain cell process signals are "turned on" in brain cells, but "turned off" in muscle cells. The genes necessary to make muscle cells contract are turned on in muscle cells but turned off in other cells.

Other genes, called *inducible genes*, are expressed differently in response to different environmental conditions. For example, injecting a certain toxin into a cell may turn on genes that can respond to this threat. Genes that can assist in attacking an invading virus will be turned on in that situation. A change in hormone levels entering a cell might trigger a response from the cell by producing more of a certain gene product. Also, genes are often expressed differently depending on the cell cycle. For example, they may be expressed only certain stages of the cycle.

Other important environmental factors that cause differential reaction of genes are gender, presence of cancerous tissue, presence of diseased cells, and other such factors.

Taken all together, what we would like to be able to do is understand what genes are turned on and turned off under different conditions. Another way to say this is that we want to learn what genes are *differently expressed* from one condition to another, and how large is that change in expression. We say that a gene which is expressed more under one condition compared to a control situation is *up-regulated*, and say that it is *down-regulated* in the opposite situation.

For example, it is certainly a primary focus of the medical field to understand the mechanisms of breast cancer. If we can understand what genes are differently expressed in breast cancer cells as opposed to non-cancerous cells, that would help us better understand breast cancer, and potentially engineer treatments for it. The same statements can be made for other diseases like leukemia or Hodgkin's.

6.3 Overview of Microarray Experiments

Microarray experiments produce enormous amounts of data. We will now give an overview of microarray experiments and how they are conducted so that we have a good understanding of how all of this data is collected and what it means.

6.3.1 Primary Purpose and Key Questions

The primary purposes of microarray experiments is to do exactly what we stated above: detect and measure the different expression levels of genes across different conditions. Some examples of the different conditions we might compare are:

1. Diseased versus normal tissue
2. Different time periods (changes over time)
3. before/after injection of a hormone
4. before/after exposure to toxins
5. before/after chemotherapy in cancer cells

6. two different forms of leukemia
7. male versus female

The types of questions we would like to answer from the data we collect generally fall into one of three categories:

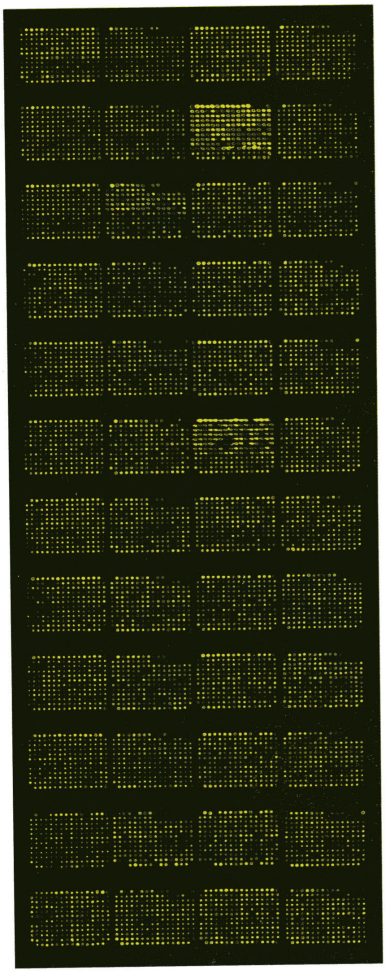
1. Which genes are differently expressed from one condition to another?
2. What genes are co-expressed across conditions (i.e., which are up- or down-regulated at the same time)?
3. Can we classify (predict) the condition under which a tissue sample was taken?

As a result, we will better understand the genes that affect and play a role in these conditions.

6.3.2 Microarray Technology

The term micro*array* is used because a microarray experiment is conducted on a solid physical medium (such as a glass slide) in which an array, or grid, is setup to collect and observe the biological data. An example of what one of these microarrays looks like is in Figure 6.1.

Figure 6.1: Example of microarray technology (Stekel, 2003)



There are two main types of microarrays. One are called *spotted microarrays*, and the other are called high-density oligonucleotide arrays. Essentially, the end result is the same whichever technology is used, in terms of the nature of the data collected and the final analyses that get conducted. However, the details of how each work and are constructed are different, as well as some of the intermediate data processing steps that need to be taken. We will focus specifically on spotted arrays as our basis for understanding microarray experiments and the data that is collected. Details of oligonucleotide arrays can easily be picked up on your own if necessary.

From Figure 6.1, we see that a spotted microarray truly resembles an array of spots. This glass slide, on which the microarray resides, can be seen to be structured in a 12 by 4 grid. Each of these 48 grids consists of a 12 by 16 array of spots, or positions. So, overall, this one microarray contains $12 \times 4 \times 12 \times 16 = 9,216$ spots. We will also refer to these as *positions* or *features*. Other spotted microarrays have a 32 by 32 array of spots within the overall 12 by 4 grid, giving a total of 49,152 positions.

To get a quick idea of the wealth of data collected during a microarray experiment, note that each position on a microarray is a point of data collection. From above, we see that a single microarray has from 9,000 to 50,000 such positions. A microarray experiment consists of a number of individual microarrays (possibly quite a few). So, we quickly wind up with lots and lots of data.

Let's talk more about each of these positions on a microarray with a quick overview. Each position represents a different gene in the organism being studied. At a particular position on the microarray is a *DNA probe* from that gene. Essentially, this is part of the DNA sequence that codes that gene. These sequences and the DNA probes can be obtained from DNA libraries. For much more information on DNA probes, and how the choice is made between bad and good ones, see Chapter 3 of Stekel (2003).

6.3.3 The Microarray Experiment

Now let's discuss the steps in the microarray experiment. There are a number of differences to this procedure as discussed here depending on the type of microarray and other factors, but the main ideas are the same.

Remember, the main idea is that we want to compare the expression levels of genes across different experimental conditions. Let's assume that we are studying a certain tissue under two conditions and review the experimental process.

We start by extracting mRNA from the tissue from each of the two conditions. Why do we extract mRNA? The mRNA present in the cells of that tissue at any point in time are representative of the genes that have actually been transcribed by the cell and are currently active. Genes that are active at that time will have some or possibly many resulting mRNA molecules present, while genes that are not active at the time will have little or no resulting mRNA present. Understanding this concept is key to understanding microarray experiments.

Each mRNA molecule is then *reverse transcribed* into *complementary DNA* or *cDNA*. This is the DNA sequence that would be complementary to the starting mRNA (i.e., T's instead of A's, C's instead of G's, etc). Why this is done we will see in a minute.

Then, an important step is that the cDNA is labelled with fluorescent dyes. The dyes are green and red. The cDNA

that came from the tissue subjected to the first condition is labelled with green, and cDNA that came from the tissue subjected to the second condition is labelled with red fluorescence. This is how we will be able to differentiate between the conditions later.

Now, these cDNA samples are placed at each position on the microarray. After certain steps are taken to ensure the integrity of the experiment, *hybridisation* is allowed to occur. The cDNA placed at a spot on the array will now hybridise with the DNA probes that had been previously placed there from a particular gene. If that gene was active in that tissue, the cDNA will hybridize to the DNA probe (since they are complimentary strands). If not, no or little hybridization will occur.

Next, the array is washed. The main idea is that any un-hybridised DNA on the array will be washed away, leaving only the hybridised DNA. It is important to notice the effect of this, and the difference we'll notice between the two conditions. If the gene was active under the first condition, that position will be green in color. If the gene was active under the second condition, that position will be red in color. If it was about equally active under both conditions, we will see yellow (a mix of red and green).

Now, we analyze the resulting image with that in mind. We will discuss this in much more detail.

6.3.4 Image Analysis of the Microarray

The raw data from the microarray experiment is contained in the resulting color image of the array (typically obtained by a scanner). A single microarray image file can be up to 32MB in size for typical experiment (Stekel, 2003) and so data storage can quickly become an issue.

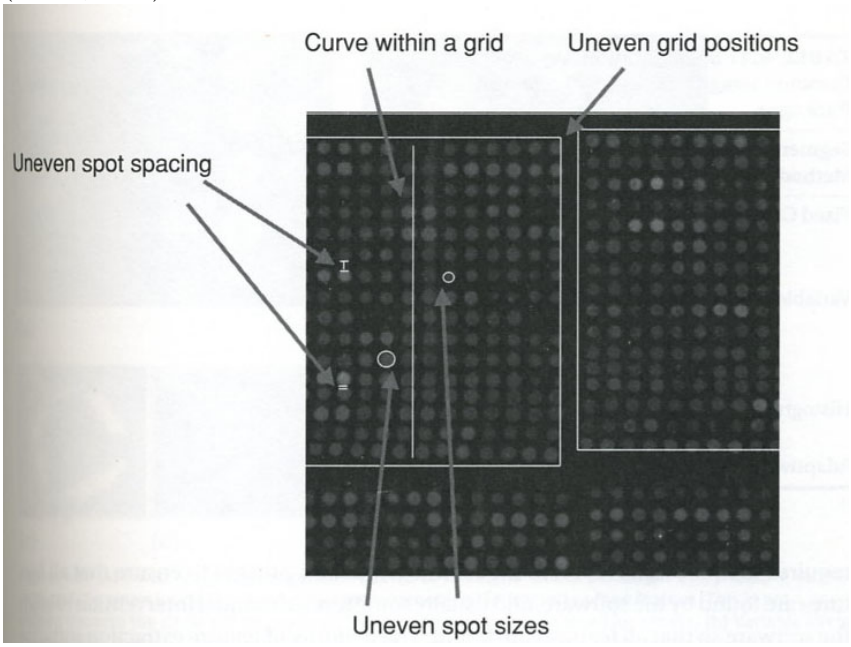
The important aspects of this image are the color and intensity of light at each position. A position (i.e., representing a gene) is comprised of a number of pixels on this image. An image processing program must determine which pixels on the image correspond to which position in the array. Even though the array is a set of seemingly regular shaped rectangular grids of positions, irregularities may be present that make this task more difficult (e.g., see Figure 6.2). The image analysis software must be calibrated in order to correctly locate the pixels associated with each position.

Once the pixels associated with a position are located, the image analysis software analyzes each position and reports a number of measurements. Results are reported for each fluorescent dye, called channels. *Channel 1* typically refers to the Cy3 green dye, and *channel 2* typically refers to the Cy5 red dye (although the definition may be switched in some cases). In other words, the color of a position is read in two pieces: how much green and how much red.

Some of the most important information that is reported for each position is:

1. The mean and median channel intensity of the pixels in that position. The maximum possible value depends on the scanner and the microarray itself, but essentially this measures "how much green" and "how much red" is in that spot. We will refer to these reported values as `Ch1Int` and `Ch2Int` for the mean intensities, and `Ch1Int(Median)` and `Ch2Int(Median)` for the medians.
2. The background mean and median channel intensity. This measures the green and red intensity levels, but for pixels nearby to the position, not in the position itself. This helps to account for the natural fluorescence of

Figure 6.2: Irregular array example, with curved grid columns and uneven grid alignment and spot spacing (Stekel, 2003).



the glass slide and other possible experimental errors (as we'll see later). We'll refer to these values as `Ch1BG` and `Ch2BG` for the background mean intensities, and `Ch1BG(Median)` and `Ch2BG(Median)` for the background medians.

3. The standard deviations of the channel intensities in the position. This measures how constant the intensity is across all pixels in the position. A high standard deviation can be a sign of a problem, such as a bright dust speck in that position which produced very high intensities only for some of the pixels.
4. A flag that reports whether the data measured for the position is good or not (typically "0" means everything seems ok).

An example of real data collected can be found on the course website in the file "17183(updated).xls". This is one array from an again experiment in humans, and represents fibroblast cells. I have cleaned it up somewhat to remove many recorded data that is unnecessary for our purposes.

6.3.5 Within Array Data Manipulation

Once data has been retrieved from the image and recorded, it must be analyzed and cleaned in a number of ways. One step is "within array data manipulation" which takes reported intensity values and standardizes them using techniques to be discussed in this section. The second major step is "between array data normalization" which normalizes data from one array to the next. That is the subject of the next section.

For data collected from a single microarray, we must control for many possible sources of experimental variation that could otherwise affect our results. This is first done by subtracting background intensities from channel intensities to get a *net intensity* for each channel. The typical calculations are:

$$\texttt{Ch1NetInt} = \texttt{Ch1Int} - \texttt{Ch1BG(Median)}$$

for channel 1, and

$$\texttt{Ch2NetInt} = \texttt{Ch2Int} - \texttt{Ch2BG(Median)}$$

for channel 2. If either of these net intensity values is not positive, the value is either reset to the lowest possible intensity value (one), or the gene is ignored. This will happen if the background intensity for the channel is higher than the intensity within the position. It could be an error or a local problem on the array, or it could represent a gene that has very low expression level on that array for that channel.

Another possible source of experimental variation that we need to account for is the different levels of green and red dye that may have been introduced into the cDNA prior to placing on the array. This could potentially skew green and red intensity levels away from the control situation. This potential bias is accounted for by computing a normalization factor for the array. The computation is based on a linear regression method which we will not discuss. For our purposes, the Normalization Factor will be reported, and lead to calculating the Channel 2 Normalized Net Intensity as follows:

$$\texttt{Ch2NormalizedNetInt} = \frac{\texttt{Ch2NetInt}}{\text{Normalization Factor}}$$

Notice that this is only computed for one of the channels (Channel 2), as this normalizes that channel to the same control level as the other. More information on the need for this normalization and the underlying regression technique can be found in Section 5.3 of Stekel(2003).

The next step is to take logs and compute a final log ratio for each position. The purpose of taking logs is to create a more even spread and bell-shaped look to the distribution of the intensity data. Otherwise, the scale of raw intensities is very right skewed, with most values in the hundreds, but a few in the 10,000 to 30,000 range. Typically log base 2 is used, and the following calculations are made:

$$\begin{aligned} G &= \log_2 \texttt{Ch1NetInt} \\ R &= \log_2 \texttt{Ch2NormalizedNetInt} \end{aligned}$$

The variable G now represents the log net intensity of the green channel for that position, and R represents the log net intensity (after normalization) of the red channel for that position.

Finally, the value typically reported for a single position on the array (and therefore, for a gene) is the difference of the R and G values:

$$L = \log_2\Big(\frac{\texttt{Ch2NormalizedNetInt}}{\texttt{Ch1NetInt}}\Big) = R - G$$

Notice that this value L for a gene is the log ratio of net intensity values for the two channels.

Remember, we are making the above calculations for each position on the array. Putting it all together, for a single array, we have $L_1, \ldots, L_g$ log ratio values, one for each of the g genes represented on the array. These values can be analyzed to give an idea as to whether a gene was up- or down-regulated across the channels on that array. Remember, the channels really represent two different experimental conditions, such as normal versus cancerous cells, or normal versus hormone-injected cells.

Our interpretation of one of these log ratios will be as follows:

L_i near 0	Gene i equally regulated across conditions
L_i near +1	Gene i up-regulated by a factor of 2 in the second condition compared to the first
L_i near -1	Gene i up-regulated by a factor of 2 in the first condition compared to the second
L_i near +2	Gene i up-regulated by a factor of 4 in the second condition compared to the first
L_i near -2	Gene i up-regulated by a factor of 4 in the first condition compared to the second
and so on.	

6.3.6 Between Array Normalization

Recall that the log ratio values computed above (and the resulting interpretations) are only for a single array. An individual array is typically used to measure the expression level of many genes for a single subject across two conditions. Of course, a single subject does not provide good information on the truly significant differences that may be present between gene expressions and conditions. To measure variability across subjects (or time periods, or other experimental units), the microarray experiment is run a number of times. If we are talking about an experiment dealing with patients, for example, the overall experiment might result in one microarray for each of 20 patients (maybe 10 have cancer, and 10 are controls who do not have that cancer). So, for multiple subjects, gene expression is measured on different arrays.

Because of the very complex processes that occur to create and extract data from a single array, there is certainly the potential for unwanted experimental variability across arrays. For example, the total amount of red and green dye may have been different from one array to the next, or maybe the scanner settings were slightly different from one scanning session to the next. We need to account and control for these possible variations between different arrays by normalizing the resulting data.

The main technique for doing this is to perform standard normalization techniques on the original G and R values from each array. For example, suppose Array 1 (call it A1) has values $G_1, \ldots, G_g$ from the green channel for each gene. Let $\bar{G}_{\text{A1}}$ be the mean of the green channel values on this array, and let $S_{G(\text{A1})}$ be the standard deviation of these values. Then, we compute a normalized G value for each gene ($i = 1, \ldots, g$) on that array as follows:

$$G'_i = \frac{G_i - \bar{G}_{\text{A1}}}{S_{G(\text{A1})}}$$

We then do the same for $R_1, \ldots, R_g$ on that array to create normalized values $R'_1, \ldots, R'_g$. Then do the same for each array, where the normalizing mean and standard deviation are computed from that array's base G and R values only.

The log ratio can now be applied to these normalized values, and the end result is a set of normalized values across arrays. The notation we will use for these log ratios is $L_{i1}, \ldots, L_{in}$, where the subscript i runs over all genes measured in the study $i = 1, \ldots, g$, and the second subscript represents the array number $(1, \ldots, n)$. Each of these L values is a log ratio as discussed before, now just computed on the between-array normalized data:

$$L_{ij} = R'_{ij} - G'_{ij}$$

6.3.7 Microarray Data on the Web

There are many microarray experiment datasets available publicly on the internet. One of the best and easiest to navigate is the Stanford Microarray Database (SMD) found at http://smd.stanford.edu/index.shtml. To get to microarray data, click Public Login on the left side. You then have many choices available in order to find an archived experiment you are interested in. Data can be viewed or downloaded, and even the microarray image itself can be viewed and interactively clicked. There is a wealth of information available here. You are encouraged to play around with some of these experiments.

6.4 Microarray Data Analysis

We will now discuss methods for answering the three key gene expression-related questions we earlier posed. The questions were

1. Which genes are differently expressed from one condition to another?
2. What genes are co-expressed across conditions (i.e., which are up- or down-regulated at the same time)?
3. Can we classify (predict) the condition under which a tissue sample was taken?

6.4.1 Identifying Differently Expressed Genes

In lecture, we will discuss three different experimental designs that are common in microarray experiments. Each leads to a slightly different view of the microarray data and the resulting log ratios. In the end, the particular design chosen leads to different options for conducting tests. For these designs, we will discuss using paired t-tests and non-pooled t-tests to identify genes that are differently expressed from one condition to the other. We will discuss these tests in class. A couple of examples are below. Data for these examples is taken from the Stanford Microarray Database, and to keep things simple, does not necessarily represent all the data available for those particular experiments (and so these do not represent definitive answers to the underlying questions). More information on these examples can be found in Perou, et al. (2000) and Golub, et al. (1999).

T-test Examples

Example 6.1: **Paired t-test example: Breast Cancer before and after chemotherapy**

Gene: T cell receptor alpha

Data: 4 patients, data has been normalized within and between arrays

$L_{1A} = 0.371$	$L_{1B} = 1.390$	$L_1 = 0.371 - 1.390 = -1.019$
$L_{2A} = 0.680$	$L_{2B} = 1.856$	$L_2 = 0.680 - 1.856 = -1.177$
$L_{3A} = 0.867$	$L_{3B} = 0.712$	$L_3 = 0.867 - 0.712 = 0.155$
$L_{4A} = 0.019$	$L_{4B} = 0.518$	$L_4 = 0.019 - 0.518 = -0.500$

The sample mean and standard deviation for $L_1, \ldots, L_4$ are

$$\bar{L} = -0.635 \qquad S_L = 0.601$$

So the test statistic is

$$T = \frac{-0.635}{0.601/\sqrt{4}} = -2.115.$$

It has a t_3 distribution, and the p-value is 0.1248. Small, but we would not reject the null hypothesis, probably due to small sample size. At this point, there is not enough evidence to suggest that this gene is differently regulated in breast cancer cells from before to after chemotherapy.

■

Example 6.2: **Non-pooled t-test example: ALL vs AML leukemia patients**

Data: 5 ALL patients (condition 1) and 3 AML patients (condition 2). Data has been normalized within and between arrays.

ALL	$L_{11} = 1.683,\ L_{12} = 1.700,\ L_{13} = 0.864,\ L_{14} = 1.907,\ L_{15} = 0.676$
AML	$L_{21} = 1.557,\ L_{22} = 1.839,\ L_{23} = 0.357$

From these we calculate

$$\bar{L_1} = 1.366 \qquad S_{L_1} = 0.555$$
$$\bar{L_2} = 1.251 \qquad S_{L_2} = 0.787$$

giving us the non-pooled t-test test statistic

$$T = \frac{1.366 - 1.251}{\sqrt{\frac{(0.555)^2}{5} + \frac{(0.787)^2}{3}}} = 0.222$$

This has approximately a t_6 distribution, producing a p-value 0.831. This is very high and we would not reject the null hypothesis, suggesting that this gene is not differently regulated in the different types of leukemia.

■

Multiple Testing Issue

Remember, in a real microarray experiment, we would conduct tests like above for all genes represented on the array (thousands to tens of thousands). We will compute a test statistic and want to make a rejection/non-rejection decision for each. But, since we are conducting so many tests, we are likely to have many false positives if we use the

standard approach to testing each hypothesis.

For example, say there are $g = 10000$ genes on the array. We conduct tests like the t-tests above on each using an individual significance level for each test of $\alpha_I = .05$. Even if all 10,000 genes were truly NOT differently expressed, our analysis would detect 500 differently expressed genes. That is, we would have 500 false positives. This problem is coming about because we are conducting so many hypothesis tests.

Recall the standard chart that is used to understand possible errors in hypothesis testing:

	Truth	
	Gene is not diff. exp.	Gene is diff. exp.
Declare it not diff. exp.	$\star$ (1)	False Negative (2)
Declare it is diff. exp.	False Positive (3)	$\star$ (4)

Historically, a common solution to the problem was to set α_I much lower in order to control the *family-wise error rate* or *FWER* at $\alpha = .05$. This approach attempts to minimize the probability of making even a single false positive across the many tests. This leads to a very small individual significance level α_I for each test (in fact $\alpha_I = .05/10000 = .000005$ in the case of 10,000 genes) , and therefore very few rejections. What happens is that we have set the false positive rate so low that the false negative rate becomes quite high. So, while we have a very small probability of declaring a gene significant when it wasn't, we have a very high probability of not declaring a gene significant when it is. We have traded off false positives for false negatives, which might not be very helpful in microarray analyses. Our goal is to find genes that are differently expressed, and we are probably willing to have some amount of false positives in order to correctly find such genes.

A new approach to multiple testing controls the *false discovery rate* or *FDR*. FDR is defined as the percentage of incorrectly rejected hypotheses divided by the total number of rejected hypotheses. In the table above, it would be the number of times we are in box (3) divided by the total number in (3) and (4). So, an FDR of .05 says that for every 100 times we reject the hypothesis (i.e., for every 100 genes we declare to be differently expressed), 5 of them will have been incorrectly rejected (i.e., 5 of those 100 genes would not truly have been differently expressed).

To use the FDR approach and what is called the *step-up* method, we decide on the false discovery rate we want to control and call it q. We then take all of our p-values from the set of tests we are conducting, rank them, and start with the smallest. Refer to this ordered ranking of p-values $P_{(1)}, \ldots, P_{(i)}, \ldots, P_{(g)}$, where again g is the total number of genes. The i-th ordered p-value $P_{(i)}$ is compared to the critical value $\frac{qi}{g}$. We search for the largest i for which $P_{(i)} \leq \frac{qi}{g}$. Let $m = \max\{i : P_{(i)} \leq \frac{qi}{g}\}$ be this largest value of i. We then reject all hypotheses related to genes $1, \ldots, m$ as labeled in the ordered ranking.

An example of performing this step-up procedure is below. The total number of genes was $g = 200$, and the FDR was set at $q = .05$. The spreadsheet shows just the first 27 ranked genes (i.e., 27 lowest p-values). In this simple example, just the first gene has a p-value that meets the criteria, so we only claim this one gene to be differentially expressed.

p-value	GeneID	Rank	qi/g	pval<=qi/g?
4.192E-05	42	1	0.00025	YES
0.00262722	180	2	0.00050	NO
0.00913572	43	3	0.00075	NO
0.01091122	156	4	0.00100	NO
0.01403091	67	5	0.00125	NO
0.01801436	88	6	0.00150	NO
0.03444255	118	7	0.00175	NO
0.04043514	117	8	0.00200	NO
0.04344267	159	9	0.00225	NO
0.04601275	34	10	0.00250	NO
0.06588947	150	11	0.00275	NO
0.06786468	175	12	0.00300	NO
0.07169765	48	13	0.00325	NO
0.09635662	16	14	0.00350	NO
0.09674936	197	15	0.00375	NO
0.10636298	66	16	0.00400	NO
0.10751475	104	17	0.00425	NO
0.11287294	44	18	0.00450	NO
0.1150526	152	19	0.00475	NO
0.11598416	196	20	0.00500	NO
0.11915675	80	21	0.00525	NO
0.12003339	12	22	0.00550	NO
0.12080788	24	23	0.00575	NO
0.12277004	28	24	0.00600	NO
0.12479633	23	25	0.00625	NO
0.12677134	148	26	0.00650	NO
0.13687688	158	27	0.00675	NO

There are other methods for using FDR to conduct many multiple comparisons as well. See Lee (2004) for a good discussion of other methods.

Other Designs and Tests

It should be noted that there are many other statistical procedures for conducting tests regarding individual genes in the microarray. We have discussed the paired and non-pooled t-tests because they are fairly simple to use and also very common. The drawback of these t-tests are that the log-ratio data need to follow normal distributions and have no major outliers.

There are plenty of alternatives to t-tests if we wish to try other methods. Some are non-parametric tests that base testing on ranks instead of t-statistics and bootstrap analysis which is more computationally intensive.

Also, we have presented just three different kinds of simple experimental designs that are common in microarray experiments. However, there are other design possibilities that are more complex and require different analysis techniques. Some examples:

1. Gene expression measured at 10 time points after injection of a hormone.
2. Compare four different types of cancer cells.

Such experiments are also common and are analyzed using techniques like analysis of variance (ANOVA) and time series methods. If you are interested, Chapters 9 and 10 in Lee (2004) give a good discussion of more complex designs of microarray experiments and methods for analysis.

6.4.2 Identifying Genes with Patterns of Co-expression

We now turn to the problem of searching for genes that are co-expressed (or co-regulated) across conditions. This can help to understand the complex interactions that genes have and the underlying gene networks that control cellular

functions or are associated with diseases. Answering this question comes down to determining which genes tend to be always turned on at the same time for a certain condition. A couple of examples are below.

1. The two conditions may be normal cells versus B-cell Lymphoma cells (a type of cancer in the lymph node system). What genes are co-expressed in these cancer cells?
2. What genes are co-regulated over time after injection of a hormone into cells?

With respect to the first of these examples, a helpful view of the data collected from the microarray experiment is as shown below.

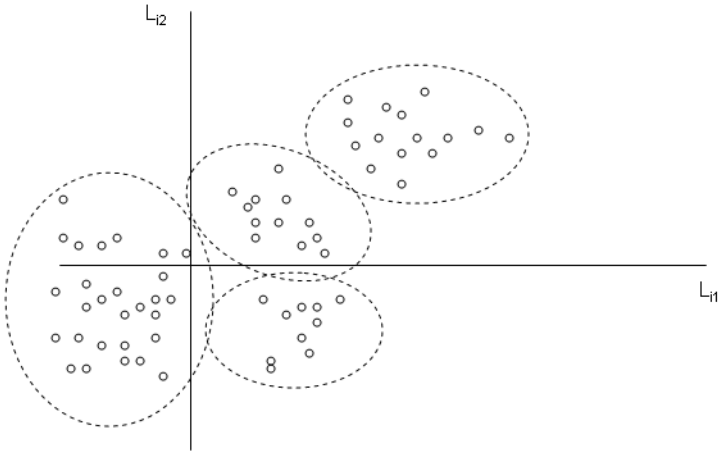
	Array 1 Subject 1	Array 2 Subject 2	$\cdots$	Array n Subject n
Gene 1	L_{11}	L_{12}	$\cdots$	L_{1n}
Gene 2	L_{21}	L_{22}	$\cdots$	L_{2n}
$\vdots$				
Gene g	L_{g1}	L_{g2}	$\cdots$	L_{gn}

The values L_{ij} in this matrix are just as we have discussed before for microarray data - it is the $\log_2$ likelihood ratio of the expression of one condition (e.g., cancer cells) compared to the second condition (e.g., normal cells) for gene i in subject j.

Now, the goal is to search for similar patterns of gene expression across subjects. In analytic terms, we want to cluster (or group) genes together that appear to have a similar expression profile (as measured by the L_{ij} across subjects.

The methods we use to do this are called clustering methods. The basic concept is that we will calculate a *distance metric* between each pair of genes (that measures how similarly or dissimilarly they are expressed) and then use clustering methods to see which genes are "close to" each other according to this distance.

As motivation for understanding how clustering methods work, consider a simple situation where there are only two subjects, so we have expression measurements L_{i1} and L_{i2} for each gene $i = 1, \ldots, g$. In this case we can make a two-dimensional scatterplot of the L_{i2} versus L_{i1} values, with a different point on the plot for each gene i. An example of such a plot is below.



In this two-dimensional example, we can think of distances between genes as the straight line Euclidean distance between points on the plot. The distance will be small the genes were expressed similarly in the two subjects, and will be large if they were expressed very differently in the two subjects. As the dotted ovals in the plot show, we can then look for groups or clusters of genes that have a similar expression across the two subjects (i.e., are close to each other). We might conclude in this small example that the genes represented within each of the oval clusters are similarly expressed across the conditions.

We need to discuss two further aspects of this problem. First, having just two subjects is nice for drawing scatterplots, but not realistic. So, we'll have to calculate distance between genes regardless of how many subjects there are. Second, we need to discuss how to more rigorously perform the clustering (i.e., draw the ovals).

Distance Metrics

In the more general setting of having n subjects, we can define distance using the standard extension of Euclidean distance to n dimensions. Specifically, for any two genes with indices a and b, we define the Euclidean distance between them to be

$$d_{ab} = \sqrt{\sum_{j=1}^{n}(L_{aj} - L_{bj})^2}$$

where the sum is over the n subjects, and this calculation can be made for any $a \neq b = 1, \ldots, g$.

A second common method for measuring the distance between two genes is based on correlation coefficients. The correlation coefficient between two genes measures the strength of their linear correlation across subjects. For two genes a and b it is defined as

$$r_{ab} = \frac{\sum_{j=1}^{n} L_{aj} L_{bj} - n\bar{L}_a \bar{L}_b}{\left(\sum_{j=1}^{n} L_{aj}^2 - n\bar{L}_a^2\right)\left(\sum_{j=1}^{n} L_{bj}^2 - n\bar{L}_b^2\right)}$$

where $\bar{L}_a$ and $\bar{L}_b$ are the means across subjects for genes a and b, respectively. The above formula is actually called the Pearson's correlation coefficient, and this calculation is very commonly provided by most analysis software programs so we do not need to worry about it in too much detail. The main thing to understand if you haven't seen correlation coefficients before is that the closer the value is to $+1$, the more positively correlated the two genes are, the closer to -1, the more negatively correlated the two genes are, and the closer to 0, the two genes are mostly uncorrelated.

Another common method for calculating correlation is called Spearman's correlation coefficient which bases the calculation on the ranks of the data instead of the data itself. Otherwise, the equation above is used as stated.

Actually, the correlation coefficient does not work correctly as a measure of distance, because it is measuring similarity between genes. So, to transform it to a distance measure, we do the following

$$d_{ab} = 1 - r_{ab}^2.$$

There are other ways to measure distance as well, but these are fairly common in co-expression analysis of microarray data.

Clustering Methods

Now that we know how to compute a distance between each gene as a measure of how differently or similarly expressed it is across subjects, we can discuss how to group genes together which have similar expression profiles. These are called clustering methods. There are a number of different clustering techniques, but we will focus on one called *hierarchical clustering*. The result of the clustering is a picture called a dendogram which puts genes near each other if they are similarly expressed, and further apart if they are not similarly expressed.

The basic algorithm for hierarchical clustering is:

1. Collect all the distances between all pairs of genes into a distance matrix.
2. Group together the two genes that are closest (smallest distance) in this matrix.
3. Recompute the distance matrix after combining the newly grouped genes into one. Now, return to Step 2 and continue until all genes have been grouped.

The recomputing of the distance matrix in Step 3 above can be done in a number of ways. The most common is called *average linkage* which just takes the average of the pairwise distances from one group of genes to the other.

At each Step 2 where we group together genes or clusters of genes, we make a new notation in our dendogram as well to demonstrate the new grouping. These techniques will all be demonstrated in class using the very good examples from Stekel (2003).

Hierarchical clustering is perhaps the most common clustering technique used in microarray analysis, but there are other clustering techniques that are used as well. Of these, *k-means clustering* and *self-organized maps* are fairly popular and discussed in the Stekel reference.

6.4.3 Classification of New Samples

The third and final important application of microarray data is its use in helping to classify new samples. In other words, microarray data has become the basis of useful and important diagnostic tools to determine whether an individual has a certain disease. For example, do they have cancer (colon, breast, lung, prostate, etc.), or which type of leukemia do they have?

The idea behind using microarray data as a diagnostic tool is the following. In our experiment which was based on a certain small number of test subjects, we analyzed two conditions (cancer cells vs. non-cancerous cells, for example). When we analyze the resulting microarray data, it may be that a certain pattern of gene expression for one, two, or many genes is very typical of cancerous cells, but extremely atypical for non-cancerous cells. This gene expression profile, then, can serve as a way to distinguish between the two.

Then, consider a single individual in the future for whom we do not know whether they have this cancer or not. We can measure this gene expression profile on that person from a microarray study, and then analyze it to determine if it points in the direction of cancer or not based on the patterns we saw from the prior study. We then *classify* the cells extracted from this individual as cancerous or not.

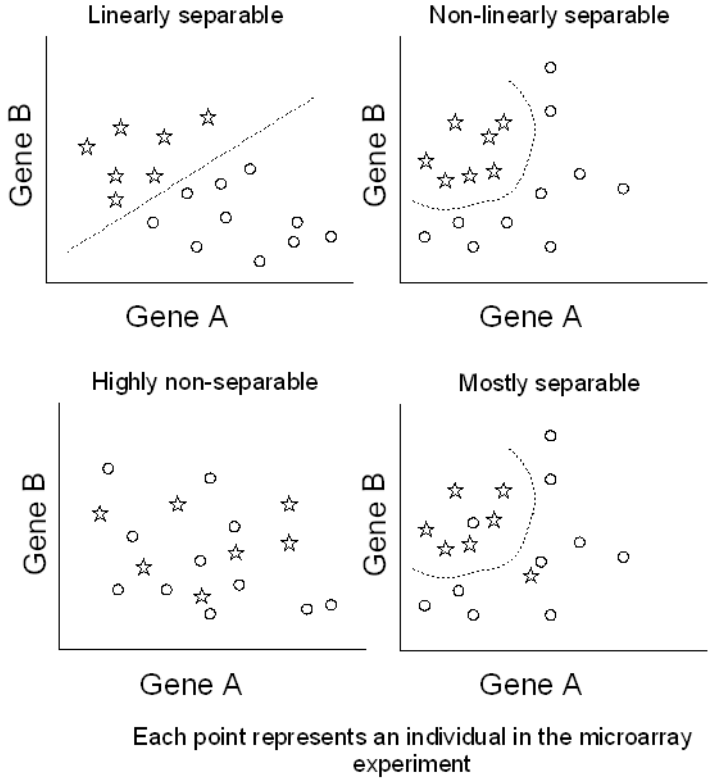
In our discussion, we will keep things fairly simple and assume that the classification we are making contains only two possible categories (for example, cancer or not-cancer or ALL versus AML leukemia). Extensions allow for situations with three or more possible categories (for example, all four forms of leukemia). We will refer to one category as category 1 and represent it with circles ($\circ$) in pictures, and the other category as category 2 and represent it with stars ($\star$) in pictures.

As a basic way of understanding how we will go about looking for a method of classification, let's draw a picture of the data for two genes at a time. For any such pair of genes, the resulting picture can be classified into one of three general categories: linearly separable, non-linearly separable, and non-separable (represented by two subcategories) as noted in Figure 6.3.
Of course, we will generally be in the non-separable case for real data, but what we will look for is a pair of genes that does the best job of separating the two classes. For example, the right bottom panel in the bottom of the figure is better than the left bottom panel. Also, there is no particular reason we need to focus on just two genes at a time. We could analyze the patterns if we looked at the resulting classification data for three, four, or any number of genes at once. But for two genes, we can at least plot the data as in Figure 6.3 and easily view the strength of the classification.

Notice the purpose of the dotted lines on the figures. We can define the line in each picture as a *classifier*. Individuals on one side of it are classified into one of the two categories, and individuals on the other side are classified into the other category. Sometimes, mistakes are made (such as in the bottom right panel). These are called *classification errors* and the percentage of such mistakes is called the *classification error rate*. The objective now is to find a function that best separates the two classes in these pictures, or in higher dimensions, where "best" is defined as having the smallest classification error rate.

Note that the functional form of the classifier need not be simple. In the upper left panel of Figure 6.3 it is just a straight line. In the upper and lower right panels, it is a fairly simple curved line. However, in general, it can be highly complex and non-linear. For example, see Figure 6.4 where the data from the lower panels of Figure 6.3 are replicated, but complex non-linear classifiers are defined.

Figure 6.3: Four possible scenarios for a particular pair of genes.



These classifiers have no classification error, but still may be not preferred for two reasons. One is that they are highly complex, and we typically want to find classifiers that are a little easier to understand and define. The second is that the apparent zero classification error rate is probably very misleading. If we were to take another sample of different subjects and classify them according to these functions found with the original subjects, we would certainly not have zero classification error (especially in the left panel situation). This situation has low, or no, error on the original dataset, but probably has high *generalization error* on new data we come across. It is really the generalization error that we want to minimize since we will be making classifications for new patients, based on the patterns we find among the original experimental subjects whose status we already know.

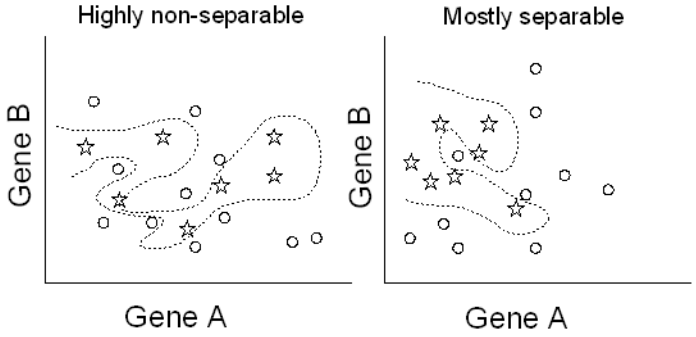
Based on the above discussion, the problem of finding a classifier has two parts: finding the functional form for the classifier, and measuring its generalization error rate.

Methods for Finding Classifiers

People in the fields of statistics, data mining, and machine learning have developed many models and techniques for performing classification going back many years, mostly not at all specific to microarray data. All of these techniques are potentially useful for microarray classification, and we will briefly mention a few, and discuss further details of one or two. Another technique called *top scoring pair* has recently been developed specifically for microarray data and will be discussed as well.

Below is a listing of techniques commonly used in these binary classification problems. We will discuss the basics of

Figure 6.4: Highly non-linear classifiers.



a few of them in class.

1. Logistic regression
2. Support vector machines
3. Decision trees
4. Nearest neighbor classifiers
5. Neural networks
6. Linear discriminant analysis

Another recently developed (here at Hopkins - see Geman, et al., 2004), and very simple methodology, is called *Top Scoring Pair* or *TSP*. The basic idea of this method is that it attempts to find a single pair of genes from the microarray study that does the best job of discriminating between the two classes. To find this best discriminating, or top scoring, pair, we calculate a score for each pair of genes i and j ($i \neq j = 1, \ldots, g$). This score is calculated based on the observed rankings of the two genes across the samples. The formula is:

$$\Delta_{ij} = \Big| \; P(X_i \leq X_j \mid \text{Class 1}) - P(X_i \leq X_j \mid \text{Class 2}) \; \Big|$$

where the X's are gene expression measurements (e.g., log likelihood ratios of R to G scores) from the microarray for each gene. For example, consider the following summary tables for the pair of genes (1,2) and (3,4) based on a sample of size 50.

	$X_1 \leq X_2$	$X_1 > X_2$	
Class 1	27	3	30
Class 2	2	18	20

	$X_3 \leq X_4$	$X_3 > X_4$	
Class 1	14	16	30
Class 2	11	9	20

We would calculate scores Δ_{12} and Δ_{34} as

$$\begin{aligned} \Delta_{12} &= \Big|27/30 - 2/20\Big| = .800 \\ \Delta_{34} &= \Big|14/30 - 11/20\Big| = .083 \end{aligned}$$

Of these two pairs, we see that the combination of genes 1 and 2 provide the best discrimination between the two classes, and therefore the highest score. We then calculate this score for all $\binom{g}{2}$ possible pairs, and pick the pair with the highest score as the basis for making predictions. It is possible that several pairs of genes will tie for highest score, in which case we use all such pairs.

Assuming only one pair has the highest score, the method for making a prediction for a future subject is fairly obvious. Say that genes i and j were to TSP. From the microarray, we measure the expression levels of these genes, X_i and X_j, for this new individual. Then, we classify the individual into Class 1 if $X_i \leq X_j$ and classify them into Class 2 if $X_i > X_j$. Notice how this serves as a very simple diagnostic test for cancer or other diseases. For example, let's say we have a patient and we want to determine (i.e., predict) if they have prostate cancer (and we label having cancer as Class 1 and not having it as Class 2). We perform a microarray study on the patient and measure expression levels $X_i = 2.6$ and $X_j = 1.1$ for these two genes that were our TSP. We then are fairly confident that this patient does not have prostate cancer according to our TSP model.

There are a number of advantages of this TSP method. First, it is very simple to derive and to understand. It can lead to a simple understanding of the biological significance of the role individual genes play in distinguishing between the classes. Other modeling methods such as logistic regression or support vector machines model very complex non-linear relationships between the expression levels of many genes at a time, making them very difficult to interpret.

A second important advantage, more from a statistical point of view, is that there are no assumptions necessary for this method. Since it is purely based on ranking the expression levels of genes (i.e., counting how many times one gene's level was higher than the other's), there are no assumptions about normality or anything else necessary. Also, this means that between-array normalization is unnecessary since all this does is scale values down, but does not change their ordering.

Finally, it is based on simple probability concepts, so it is also easy to understand and analyze from a probabilistic viewpoint. It turns out that this methodology is about as accurate, or in some cases more accurate, than much more complex methods dependent on many genes at once.

Methods for Estimating the Accuracy of a Classifier

We have mentioned that estimating the accuracy of a classifier is very important, and comes down to estimating what we called its generalization error. In other words, how good will the resulting predictions be if I begin using to make predictions for new, previously unseen, subjects? If you think carefully about this question, it might seem impossible to measure generalization error, because by definition, we have do not yet have data for these new, previously unseen subjects, so how can we determine how good the predictions will be? This is very true, so the answer is that we will have to make judicious use of the sample of size n that we do have in order to calculate the classifier's generalization error rate.

There are a few ways to do this. The two most common are called the *holdout method* and *cross-validation*. The second of these, cross-validation, is much more useful when there are small sample sizes, which is generally the case with microarray data. So, we will focus on this method and describe it more fully in lecture.

Bibliography

[1] Weir, B.S. (1996) *Genetic Data Analysis II.*

[2] Stekel, D. (2003) *Microarray Bioinformatics.*

[3] Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.*

[4] Perou, C.M., et al. (2000) Molecular portraits of human breast tumours. *Nature* 406:747-52.

[5] Golub, T.R., et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286:531-37.

[6] Sherlock, G., et al. (2001) The Stanford Microarray Database. *Nucleic Acids Research* 29:152-55.

[7] Ewens, W., Grant, G. (2001) *Statistical Methods in Bioinformatics.*

[8] Lange, K. (2002) *Mathematical and Statistical Methods for Genetic Analysis, 2nd ed.*

[9] Lee, M.T. (2004) *Analysis of Microarray Gene Expression Data.*

[10] Evett, I., Weir, B.S. (1998) *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists.*

[11] Hartl, Daniel L. and Elizabeth W. Jones (2005) *Genetics: Analysis of Genes and Genomes, 6th ed.*

[12] Elrod, Susan and William Stansfield (2002) *Schaum's Outlines: Genetics, 4th ed.*

[13] Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48: 443-53.

[14] Gotoh, O. (1982) An improved algorithm for matching biological sequences. *Journal of Molecular Biology* 162: 705-08.

[15] Mount, David W. (2004) *Bioinformatics: Sequence and Genome Analysis, 2nd edition.*

[16] Taylor, Howard M. and Samuel Karlin (1984) *An Introduction to Stochastic Modeling.*

[17] Orengo, C.A., Jones, D.T, and J.M. Thornton (2003) *Bioinformatics: Genes, Proteins, & Computers.*

[18] Zhang, M.Q. (1998) Statistical features of human exons and their flanking regions. *Human Molecular Genetics* 7(5): 919-32.

[19] Snyder, E. E. and Stormo, G. D. J. (1995) Identification of coding regions in genomic DNA. *Journal of Molecular Biology* 248: 1-18.

[20] Hertz, G.Z, and G.D. Stormo (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 15:563-77

[21] Eisenberg, E. and E.Y. Levanon (2003) Human housekeeping genes are compact. *Trends in Genetics* 19:362-65.

[22] Lee, Mei-Ling Ting (2004) *Analysis of Microarray Gene Expression Data.*

[23] Geman, D., d'Avignon, C., Naiman, D., and Winslow, R. (2004) Classifying gene expression profiles from pairwise mRNA comparisons. *Statist. Appl. in Genetics and Molecular Biology* 3.

[24] L. Xu, A-C Tan, D. Naiman, D. Geman and R. Winslow (2005) Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. *Bioinformatics* 21:3905-3911.

[25] A-C Tan, D. Naiman, L. Xu, R. Winslow and D. Geman (2005) Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics* 21: 3896-3904.