Chapter 1

Background

Probabilistic models in biology are applied in many and diverse problems. In this course, we will focus on applications to genetics and to biological sequence analysis. It is assumed that the reader has a general familiarity of the current understanding of heredity—that heredity is carried in genes encoded in the DNA of each cell and passed to future organisms in the process of reproduction. Section 1 of this chapter, on genes and DNA, is a synopsis of enough background biology to understand the models presented and analyzed in these notes. There is also an html version of this section available on the course web site, with informative links.

It is assumed as well in these notes that the reader has taken a first course in probability. We will be building on this knowledge and introducing further concepts in statistics and probability. Section 2 of this chapter reviews some key points of elementary probability, to establish notation and single out basic and often-used concepts. A full, but concise, review of definitions and formula from beginning probability is provided for reference in the Appendix.

1.1 Genes and DNA

1.1.1 Mendelian genetics and its molecular basis

Prior to the amazing advances of molecular biology in the second half of the twentieth century, heredity was understood in terms of Mendelian genetics. Using pollination by hand, Mendel crossed pea plants and observed the relationship of their traits to those of their offspring. He observed, for example,

that the peas themselves were either wrinkled yellow, wrinkled green, smooth yellow, or smooth green, and he counted the number of individuals with each trait through several generations of crossings. To explain his results, Mendel postulated first that traits are passed down from generation to generation in discrete units, which we now call **genes**. Thus, for peas, he proposed that there is a gene for pea color, with two variants, green and yellow, and a second gene for skin type, again with two variants, smooth and wrinkled. The different variants governing one trait or characteristic are now called **alleles** of the gene; thus the gene for pea color is said to have two alleles, yellow and green. Mendel also postulated that each organism contains two copies of each gene, that if the two copies contain different alleles, only one allele of the two, the dominant allele, is expressed physically in the organism, and that genes for different traits are passed down independently of one another. This last postulate—the independent assortment postulate–means that for a specific gene, offspring can have any combination of the alleles present in the parents, as long as one comes from the father and one from the mother, and these combinations are statistically unlinked among the offspring.

Mendel's postulates are not all strictly correct, but they provide the right framework for genetic analysis once modified to account for how reproduction works. The seminal concept of hereditary units, genes, has turned out to be correct. The independent assortment postulate is often true, as in Mendel's pea experiment, but not universal; oftentimes assortment of alleles is statistically linked, for reasons to be explained below. Often, also, one allele is not strictly dominant. Nevertheless, with the correct modifications, Mendel's theory is the basis of modern genetics.

In early genetics, the idea of a gene was an inference from experiments; Mendel and his successors would have had little basis for speculating on its actual physical manifestation. But the theory's success suggested that genes exist as real physical entities. This is correct, and the development of genetics and molecular biology has progressively elucidated the physical basis of the gene and of how it works, down to the molecular level.

The first step was connecting genes to **chromosomes** and their role in sexual reproduction. In **eukaryotic** cells, that is, cells with nuclei, the chromosomes are large complexes of protein and nucleic acid residing in the nucleus. In **prokaryotic** (without nucleus) cells, such as those of bacteria, the chromosome is generally a circular loop of DNA. By the early twentieth century, it was understood that each gene may be physically identified with a small and specific location in the chromosome. The position along

the chromosome at which a gene occurs is called its **locus**. Alleles are then understood to be alternate forms of a gene that can reside at its locus. It was also observed that the normal body cells of most organisms are **diploid**, that is, they have two copies of each chromosome in their body cells, (with the exception of the X chromosome in males of species with two sexes). Diploidy is the physical basis of Mendel's hypothesis that each organism contains two copies of each gene; each chromosome of a pair carries a copy. Diploidy allows individuals to carry two different alleles of gene.

The manner in which chromosomes are passed on to progeny in sexual reproduction explains Mendelian assortment and when it is independent and when not. For sexual reproduction, organisms make gametes—eggs or sperm—which are **haploid** cells, meaning they contain only one copy of each chromosome. Mating causes the union of two gametes, producing a diploid zygote that develops into the mature individual. Thus offspring get half of their genetic material from each parent. Gametes are produced by a process called **meiosis**, in which the genetic material of an original diploid cell is duplicated once and then divided among four haploid progeny cells. Consider the two copies, R_1 and R_2, of a chromosome in a diploid individual. In the simplest case, these copies maintain their identity throughout meiosis; two of the final haploid gametes resulting from a single meitotic division will contain a copy of R_1, and two will contain a copy of R_2. If ℓ is a locus on this R chromosome and A denotes the allele at ℓ on R_1 and a denotes the allele at ℓ on R_2, then half of the gametes get A, and half get a. Consider copies S_1 and S_2 of a different chromosome of the diploid parent, on which resides a locus j, and suppose that B is the allele at j on S_1, while b is the allele at j on S_2. Suppose again that S_1 and S_2 are transmitted whole in meiosis. If we follow both pairs of chromosomes in a meiotic cell division, we will see that the the different copies of the R and S chromosomes can be transmitted in any combinations, that is, the resulting gametes can end up with copies of any of the four combinations (R_1, S_1), (R_1, S_2), (R_2, S_1), and (R_2, S_2). Furthermore, if we record the results of many meiotic divisions, we will find that these different combinations occur with equal probability, that is, the copies of the R and S chromosome assort themselves independently. The alleles on loci ℓ and j go along for the ride. Thus, the gametes can contain any of the allele pairs, (A, B), (A, b), (a, B), or (a, b) corresponing to the four different combinations of chromosome copies, and the transmission of A or a is independent of the transmission of B or b. This explains the independent assortment of traits in Mendel's experiment; it is due to having

genes that reside on different chromosomes.

The story for two loci on the same chromosome is different. Suppose that ℓ and k are loci on the R chromosome, and suppose that copy R_1 has allele A at locus ℓ and allele C at locus k, while R_2 has allele a at ℓ and allele c at k, In the scenario in which R_1 and R_2 are transmitted whole to the gametes, there can be no assortment of the alleles at ℓ and k. Each gamete will contain either A and C or a and c. In actual fact, such strict linkage is never observed, and alleles at loci on the same chromosome can assort. The reason is a process in meiosis called **recombination**, in which pieces pieces of R_1 and R_2 exchange in meiosis, so that the copies of R in the gametes are an amalgams of material from R_1 and R_2. Recombination can thus separate loci on the same chromosome of the parent into different gametes. Figure 1 is a schematic explanation. On the top are the two original copies R_1 and R_2 of a chromosome. On the bottom are the results of a recombination in meiosis. A and a, and C and c are the alleles at two loci.

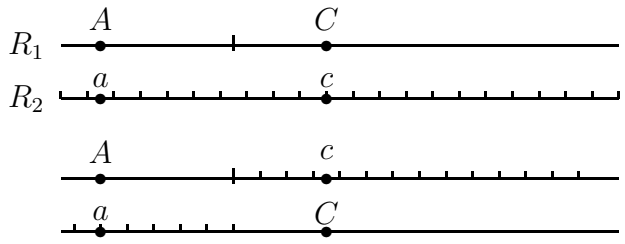

Figure 1.

Recombinations and where they are located seem to be random events. They can occur throughout the chromosome and may or may not occur between two given loci. Thus, while alleles at loci on the same chromosome can assort by recombination, they are not transmitted independently, because with some probability no recombination occurs, causing them to be linked.

The discoveries about genes and chromosomes led to a refined definition of gene. We quote here from the glossary of W.J. Ewens' monograph, **Population Genetics**, Methuen, London (1969): a gene is "a minute zone of a chromosome which is the fundamental unit of heredity. A gene partially or wholly governs the expression of a certain character or characters in an individual." This definition still does not take us down to the molecular level, nor explain how inheritance is carried on the chromosome, but it is adequate to understand the modern theory of population genetics.

In the last half of the twentieth century, an understanding of the gene at the molecular level emerged. Here are the main points, necessarily oversimplified and over-generalized—biology is full of exceptions to the rule! You will need to understand these points to appreciate the models and problems arising in biological sequence analysis.

1. DNA (RNA in viruses) is the heredity-carrying material of the chromosome.

2. DNA is a linear, unbranched polymer consisting of two complementary chains of nucleotides; each nucleotide consists of one of the four nucleic acids, adenine or guanine (the pyridines), or cytosine or thymine (the pyrimidines), attached to a deoxyribose sugar molecule, and a phosphate group, also attached to the sugar. The nucleotides of one chain link to each other through bonds between the phosphate groups. In a single-stranded chain, the sugar molecules and phosphate groups thus form a backbone supporting a sequence of nucleic acids. The complementary chains in turn link to one another by hydrogen bonds between the nucleic acid, with adenine bonding only to thymine and guanine only to cytosine, and the chains twist around one another to form the famous double helix. A unit consisting of paired nuceotides is called a **base pair**. It is abbreviated by *bp* and used as a unit of length when discussing DNA. A unit of 1000 base pairs, a kilobase pair, is abbreviated *kb*.

   The deoxyribose sugar molecules has 5 carbon atoms, which are labelled from 1' to 5'. In the DNA backbone chain, the 3'-carbon of a deoxyribose molecule links to the 5'-carbon of an adjacent deoxyribose by way of a phosphate group. This 5' to 3' linkage gives an orientation to a single strand of DNA. Hence, ignoring the details of molecular structure, one can represent a DNA strand abstractly as an ordered sequence of nucleic acids, and this sequence carries much of the information needed to do genetic analysis. As an example, consider the following representation of a made-up DNA sequence:

$$5' - AGCGTTAGGCTATTAGGCGA - 3'$$

   The letters A, T, C, and G are shorthand for the 4 bases. In the topmost strand, the leftmost base is A (adenine); if it links to any base on the left, it does so via the 5' carbon of the deoxyribose (not shown)

of the nucleotide. The 3'-carbon of this first nuceotide then links to the 5'-carbon of the deoxyribose of the next nucleotide, to which base C is affixed, and so on down the line. Of course in the cell, DNA is found usually as a double stranded molecule. If the example just given were part of one of those strands, the double stranded piece to which it corresponded could be represented as:

$$5' - AGCGTTAGGCTATTAGGCGA - 3'$$
$$3' - TCGCAATCCGATAATCCGCT - 5'$$

The top and bottom strands are linked by hydrogen bonds, as previously mentioned, with links only between A's and T's or between G's and C's, and the complementary (bottom) strand is oriented in the complementary, $3' - 5'$, direction. Therefore, representation of a single strand is enough to determine the base pairs of the double-stranded molecule.

Important note: When dealing with single stranded DNA, we shall take the 5'-3' direction as the default orientation and usually omit its specification.

3. Chromosomal DNA holds the instructions for the production of **polypeptides**, which are the linear chains of amino acids making up proteins. In essence, the hereditary material of a chromosome can be thought of as a single, long string of DNA. For each polypeptide an organism produces there is a corresponding spot along a this DNA string with instructions for constructing it. DNA stores the instructions by means of the famous **genetic code.** There are 20 different amino acids that go into the making of any protein. Different three-letter "words" or **codons** from the DNA alphabet $\{A, T, C, G\}$ either code for one of the 20 amino acids or signal a start or stop to coding. A complicated transcription mechanism in the cell, "reads" the DNA sequence and builds a protein (polypeptide) by sticking together amino acids in the order specified by the successive codons. The coding regions of a DNA sequence are called **exons**—they are the regions which are (ex)pressed. Between the exons are non-coding regions called **introns**, which, in higher animals, are quite extensive and constitute more of the DNA than the exons. The function of introns, if any, is not well understood.

The complete DNA sequence of all the chromosomes in the nuclei of an organism is called its **genome**.

This molecular-level picture leads to the most refined and precise definition of a gene, the so-called **one gene, one polypeptide theory**; a gene is a unit, identifiable with a stretch of coding DNA, that governs the production of a polypeptide. (Strictly speaking, this is not always true—there are DNA segments that get expressed as RNA only, but not as protein.) In this interpretation, the locus of a gene is that specific stretch of DNA where the code for its associated polypeptide is located. The alleles of a gene are variant DNA sequences at its locus. They will get expressed as different versions of the associated polypeptide, which, if sufficiently different in their biochemical properities, will in turn produce individuals with varying traits. Sickle-cell anemia provides a classic and dramatic illustration of the one gene, one polypeptide theory. A specific gene encodes instructions for fabrication of the beta sheet of hemoglobin. In normal hemoglobin, glutamic acid occupies the sixth position in the beta chain (you don't have to know what this is—it's just a part of the hemoglobin molecule), and it is encoded for in the gene by the DNA triple GAG. The allele responsible for sickle-cell anemia differs from the normal allele in only one base! The codon for the sixth amino acid in the beta sheet is GTG instead of GAG and it is expressed as valine rather than glutamic acid in the beta sheet. This one substitution so affects the structure of hemoglobin as to cause sickle-cell anemia.

1.1.2 Genotypes, phenotypes, and polymorphisms

The **genotype** of an individual is the genetic endowment carried by its chromosome. You can think of it as a list of the alleles an individual has at the loci of its chromosomes. The word *genotype* is actually used in two senses, one very broad, the other restricted. In the broad sense, ocurring in general discussions, it refers to an organism's full genetic composition, that is, the alleles at all its loci. The restricted usage applies to the description of alleles at a given, relatively small, set of loci, and is used in genetical studies of particular genes. Consider, for example, the study of a single locus ℓ. Each individual has two copies of the locus, one per chromosome. The genotype of an individual at ℓ is therefore represented by a pair of letters representing the two alleles that occur. on the chromosomes containing ℓ. If there are two possible alleles A and a, the genotype will be one of the pairs. AA,

Aa, or *aa*. In this context, the following terminology is very important. An individual with identical alleles at a locus (that is, *AA*, or *aa*) is said to be **homozygous at the locus**; an individual with different alleles (*Aa*) is said to be **heterozygous**.

It is important to distinguish an organism's genotype from its **phenotype**. The phenotype is the set of the organism's physical, biochemical, and behavioral traits—how it looks and functions in the world. As with the word *genotype*, the term *phenotype* can be applied either in a broad sense to the whole organism or, in a restricted sense, to a specified set of traits. The relationship between the genotype and phenotype, between genes and traits, is complicated and varied. First, the expression of an individual's genetic endowment, as it develops into a mature individual and lives its life, is mediated by the environment. Environmental influences will cause even genetically identical individuals—in other words, clones—to have different phenotypes; for example, supplied with different amounts of nutrition, clones might grow to different sizes. Second, individuals with different genotypes can have the same phenotype. This happens because of the phenomenon of allelic dominance. For example, in Mendel's pea plants, the allele for yellow color dominates that for green. When a plant is a heterozygote for color, that is, possesses one allele for green and one for yellow, it will have yellow peas, just the same as a homozygous plant with two yellow alleles. Thus heterozygotes and homozygotes for yellow will be phenotypically indistinguishable. Dominance explains the fact observed in Mendel's experiment that a cross between two plants with yellow peas can give rise to progeny with green peas, because if both parents are heterozygous, it is possible for their offspring to end up with two green alleles. One allele is not always dominant over another. They may both affect the phenotype, leading to a situation in which heterozygotes and homozygotes do differ in phenotype. A third complication in the relation between genotype and phenotype is the complex way in which genes, which code for proteins, affect traits. There are of course traits controlled by a single gene, such as in the example of sickle cell anemia, described above. The pea traits studied by Mendel also appear to be controlled by single genes—a very fortunate fact for Mendel as it led him to the idea of discrete hereditary units. But other traits may be governed by the interaction of many genes, so that various combinations of alleles can cause subtle difference between individuals in the traits' many variables. These are called **quantitative traits**, and it is in general difficult to sort out their genetics.

Let us now step back from a focus on individuals and consider populations. Fix a particular locus or gene to consider in a population of individuals of the same species. The situation in which two or more alleles at this locus are present in the population is called a **polymorphism**. Actually this is not quite right. Technically, a polymorphism is defined to occur only if the frequency of the most common allele is less than or equal to 0.95. The idea is to restrict the term to those situations in which there is real allelic diversity that is naturally maintained, since for almost any locus in a large population there will be a least a few individuals with a variant gene, perhaps one recurrently introduced by rare mutations. In practice, it seems that geneticists loosen the 95% rule when discussing rare genetic diseases.

At the molecular level, a polymorphism in a gene is due to variation, a polymorphism, in the DNA sequence encoding the gene. Genes reside in exons, so genetic polymorphisms are by definition based on DNA polymorphism in exons. But there are also sites in introns, called **DNA markers**, at which variation of the DNA sequence occurs across populations. The term *polymorphism* is used to describe variation in DNA markers, as well.

Polymorphisms in the molecular structure of DNA take a number of forms. A **single nucleotide polymorphism**, abbreviated SNP, is a polymorphism in the nucleic acids present at a specific base pair. Sickle-cell anemia is an example of a SNP, because, as explained, it is caused by a substitution at a specific base pair in the gene coding for the beta sheet of hemoglobin. Polymorphisms occur also at satellite DNA sites. DNA satellites are locations in which a short DNA word is repeated over and over, as in $CACACACACACACACACACACA$, a tandem repeat. (The word *satellite* comes from the term *satellite bands* used to describe the bands by which repeats are revealed in centrifugal fractioning of DNA.) Satellites come as *microsatellites*, up to 20kb long with repeat units up to 25bp, and *minisatellites*, less than 150bp with small repeat units, typically tandem repeats. They can be highly polymorphic in the number of repeat units, and so they serve as useful DNA markers. A third type of molecular polymorphism is the **restriction frament length polymorphism**, or RFLP. This will be described in more detail in a later chapter. For now we give only a brief explanation. There is a class of enzymes, called restriction enzymes, which cut DNA at specific sequences of bases called recognition sequences. For example, the recognition sequence of the *Alu*1 enzyme is

$$5' - AGCT - 3'$$

$$3' - TCGA - 5'$$

and when it encounters this sequence it cuts through the DNA between the second and third base pairs. If a DNA segment is mixed with the enzyme it will be broken into relatively small fragments, the restriction fragments, in a process called a restriction sequence digest. Suppose there is a location of four base pairs that is polymorphic for the recognition sequence of *Alu*1—that is, some individuals have the recognition sequence at the locus, others don't. Then this location will be cut in some individuals but not in others, and this means that the distribution of fragment lengths in digests of their DNA will vary, thus producing an RFLP.

1.2 Probability review

It is assumed that you have seen most of this elementary material; the review should help you recollect and organize it for the applications of this course.

1.2.1 Some basic definitions

The basic definitions of probability theory really just provide a uniform framework in which to construct probability models. Take the concept of a probability space. Suppose we are studying an experiment, such as a mating trial, with random outcomes. We form a set Ω of all the possible outcomes, which in this chapter, will always be a finite set. A probability model of the experiment would just be this set Ω of outcomes together with an assignment of numbers $\mathbb{P}(A)$ in $[0, 1]$ to subsets A of Ω, where $\mathbb{P}(A)$ is to be interpreted as the probability that the outcome of a trial belongs A. As such pairs $(\Omega, \mathbb{P})$ are so fundamental an abstraction of probability modelling, probabilists dignify if with the name **probability spaces**. (The assignment $\mathbb{P}$ of course cannot of course be arbitrary; it must satisfy, in the case of finite S, at least the properies

- $\mathbb{P}(\Omega) = 1$, and;
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ if A and B are disjoint.

This means that $\mathbb{P}(A) = \sum_{s \in A} \mathbb{P}(\{s\})$, where $\mathbb{P}(\{s\})$ is the probability that precisely outcome s occurs, so, in the case of a discrete Ω, $\mathbb{P}$ is determined by the probabilities of the individual outcomes.)

You should also know another concept with which to frame a probability model— the random variable. Often, the outcome of an experiment can be represented as a real number. We label this potential outcome by a capital letter, usually from the end of the alphabet, like X, and call it a random variable. Recall that the random variable X is **discrete** if the set of possible values it can take on, is a finite or countable set of real numbers $S = \{s_1, s_2, s_3, \ldots\}$. For a such an S, a probability model for the experiment whose outcome is represented by X consists in specifying the probabilities

$$p_X(s) \stackrel{\triangle}{=} \mathbb{P}\left(X = s\right), \qquad \text{for } s \text{ in } S. \tag{1.1}$$

(The symbol $\stackrel{\triangle}{=}$ means we are defining the symbol on the left of the equation by the expression on the right.) p_X is a function on S called the **probability mass function** of X. More complicated experiment, or a series of experiments, might result in a vector $Z = (X_1, \ldots, X_n)$ of random variables. When these are all discrete, the **joint probability mass function** of $(X_1, \ldots, X_n)$ is

$$p_Z(s_1, \ldots, s_n) \stackrel{\triangle}{=} \mathbb{P}(X_1 = s_1, \ldots, X_n = s_n), \tag{1.2}$$

for all possible values $(s_1, \ldots, s_n)$ of $(X_1, \ldots, X_n)$.

We assume that the reader is familiar with these basic notions. We assume as well that the reader knows what it means for events or random variables to be independent, the definition of conditional probabilities and Bayes rule, and the definition and rules for computing expected values and variances of random variables. A brief summary of all these things may be found in the appendix. We recall here the chief definitions.

Two events A and B are independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) \tag{1.3}$$

Two random variables X and Y are independent if

$$\mathbb{P}\left(X = s, Y = t\right) = \mathbb{P}(X = s)\mathbb{P}(Y = t) \quad \text{for all possible values } s \text{ of } X \text{ and } t \text{ of } Y. \tag{1.4}$$

Generalizing, the random variables $X_1, X_2, \ldots, X_N$ are independent of one another if

$$\mathbb{P}\left(X_1 = s_1, X_2 = s_2, \ldots, X_n = s_N\right) = \mathbb{P}(X_1 = s_1)\mathbb{P}(X_2 = s_2) \cdots \mathbb{P}(X_N = s_N) \tag{1.5}$$

for all possible values $s_1, s_2, \ldots, s_N$ of $X_1, \cdots, X_N$. Finally, for a discrete random variable X, recall the definition of its expected value (or mean value), and variance:

$$E[X] \triangleq \sum_{s \in S} s p_X(s) \tag{1.6}$$

$$\mathrm{Var}(X) \triangleq E\left[(X - E[X])^2\right] = \sum_{s \in S} (s - E[X])^2 p_X(s). \tag{1.7}$$

1.2.2 Bernoulli, binomial and multinomial random variables

Bernoulli random variables A Bernoulli random variable is just a random variable that takes on only two possible values s_1 and s_2. When discussing Bernoulli random variables, we shall always assume the default values $s_1 = 0$ and $s_2 = 1$, unless we explicitly say otherwise. There is not much choice here for the probability mass function. It has to be of the form,

$$p(0) = 1 - p \qquad\qquad p(1) = p \tag{1.8}$$

where p is a number in the interval $[0, 1]$. If X is a Bernoulli random variable that has this as its probability mass function, then, by definition, p is the probability that X equals 1: $\mathbb{P}(X = 1) = p(1) = p$. It will be very useful in the future to use the following expression for the function defined in (1.8):

$$p(s) = p^s (1 - p)^{1-s} \qquad \text{for } s \text{ in the set } \{0, 1\}. \tag{1.9}$$

It is often also convenient to use q to denote $1 - p$. In this case, we would write (1.9) as $p(s) = p^s q^{1-s}$.

One of the simplest applications of the Bernoulli model is a trial that randomly results in either a *success* or a *failure*. To generate a Bernoulli random variable from such a trial assign a 1 to a success and a 0 to a failure. The language of success and failure trials is a convenient way to discuss Bernoulli random variables. We use the description "X is a Bernoulli random probability with probability of success p" to indicate a random variable with the probability mass function given in (1.8), whether or not we are thinking in terms of successes and failures.

Example 1: Of course, a Bernoulli random variable is the probability model for a coin toss. Another favorite use is to count whether a given event occurs

or not. Here is an example. Suppose we are studying a locus with two alleles A_1 and A_2 in a large population. We attach to each individual a label indicating its two alleles; the possible labels are A_1A_1, A_1A_2, and A_2A_2 and they are called the genotypes of the individuals. We perform the following experiment. Draw an individual at random and record its genotype. If the genotype is A_1A_1, let $X = 1$; otherwise, let $X = 0$. X is an *indicator* random variable for the event of drawing an A_1A_1. Suppose the experiment is repeated ten times, and let X_i indicate the outcome of trial i. Then each X_i is a Bernoulli random variable and $\sum_{i=1}^{10} X_i$ counts the total number of times the event of drawing an A_1A_1 individual occurs in the 10 trials.

The expectation and variance of the Bernoulli are easy to compute directly. If X is Bernoulli with success probability p,

$$\begin{aligned} E[X] &= 0 \cdot (1-p) + 1 \cdot p = p, && (1.10) \\ \mathrm{Var}(X) &= E[X^2] - (E[X])^2 = 0^2(1-p) + 1^2 p - p^2 = p(1-p). && (1.11) \end{aligned}$$

Binomial random variables. Let n be a positive integers and let p be a number in $[0, 1]$. A random variable Y has the **binomial distribution** with parameters n, p if Y takes values in the set of integers $\{0, 1, \ldots, n\}$ and has the probability mass function

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad \text{for } k \text{ in } \{0, 1, \ldots, n\}. \tag{1.12}$$

The binomial distribution is fundamental in applications because of the following fact. Let $X_1, X_2, \ldots, X_n$ be *independent* Bernoulli random variables each with success probability p. Then the sum $\displaystyle\sum_{i=1}^{n} X_i$ has the binomial distribution with parameters (n, p). But, as in example 1, this sum is just the total number of successes in the n trials. So, the binomial distribution is the probability distribution for the number of successes in n independent and identically distributed trials where p is the probability of success in each trial. To help you recall this result, you are asked to derive it in the exercises.

To compute the expectation of the binomial, recall these basic facts:

(i) If $Y_1, Y_2, \ldots, Y_N$ are *any* random variables, each having a finite expectation,

$$E\left[Y_1 + Y_2 + \cdots + Y_N\right] = E[Y_1] + E[Y_2] + \cdots E[Y_N]; \tag{1.13}$$

(ii) If $Y_1, Y_2, \ldots, Y_N$ are *independent* random variables, each having a finite variance,

$$\mathrm{Var}\,(Y_1 + Y_2 + \cdots + Y_N) = \mathrm{Var}(Y_1) + \mathrm{Var}(Y_2) + \cdots \mathrm{Var}(Y_N). \quad (1.14)$$

Now let $X_1, \ldots, X_n$ be independent, Bernoulli random variables, each with probability p of success, so that $Y = \sum_1^n X_i$ is binomial with parameters (n, p). Using (1.13) and (1.14) and the mean and variance of the binomial random variable,

$$E[Y] = np \qquad \mathrm{Var}(Y) = np(1-p). \quad (1.15)$$

Example 2: Sampling with replacement Consider a box with a large number of marbles, which are either red or green. Let p be the proportion of marbles in the box that are red, that is, the number of red marbles divided by the total number of marbles. If, in a random draw, any marble is equally likely to be chosen, p is the probability of drawing a red one. A single random draw is a sampling from the population of marbles. Suppose that we sample, then return the marble to the box, sample again by a random draw, return the marble, etcetera. This is called sampling with replacement. Each sample is independent of the others and the probability that red is drawn is p each time. Therefore the number of red marbles in n draws will be a binomial random variable with parameters (n, p).

Multinomial distributions We start with an example. Suppose that we have a box with three colors of marbles, red, green, and blue. Let p_1 be the probability of drawing a red, p_2 the probability of drawing a green, and p_3 the probability of drawing a blue. Of course, $p_1 + p_2 + p_3$. Sample the box n times with replacement, and let Y_1 be the number of reds drawn, Y_2 the number of greens drawn, and Y_3 the number of blue. The random variables Y_1, Y_2, Y_3 are not independent; indeed, they must satisfy $Y_1 + Y_2 + Y_3 = n$. What is their joint distribution? It is:

$$\mathbb{P}\,(Y_1 = k_1, Y_2 = k_2, Y_3 = k_3) = \frac{n!}{k_1! k_2! k_3!} p_1^{k_1} p_2^{k_2} p_3^{k_3}, \quad (1.16)$$

for any non-negative integers k_1, k_2, k_3 such that $k_1 + k_2 + k_3 = n$. To see this, consider a specific sequence of n draws that results in k_1 red, k_2 green, and k_3

blue marbles. Since the draws are independent their probabilities multiply, and so the probability of any such sequence is $p_1^{k_1} p_2^{k_2} p_3^{k_3}$. On the other hand there are a total of of $\dfrac{n!}{k_1!k_2!k_3!}$ is the different sequences of draws giving k_1 red, k_2 green, and k_3 blue marbles. Thus the total probability is given as in (1.16).

The general multinomial distribution is a generalizaiton of formula (1.16). To state it, recall the general notation,

$$\binom{n}{k_1 \cdots k_r} \triangleq \frac{n!}{k_1! \cdots k_r!}.$$

Fix two positive integers n and r with $0 < r < n$. Suppose that for each index i, $1 \leq i \leq r$, a probability p_i is given satisfying $0 < p_i < 1$, and assume also that $p_1 + \cdots + p_n = 1$. The random vector $Z = (Y_1, \cdots, Y_r)$ is said to have the multinomial distribution with parameters $(n, r, p_1, \ldots, p_r)$ if

$$\mathbb{P}\left(Y_1 = k_1, \ldots, Y_r = k_r\right) = \binom{n}{k_1 \cdots k_r} p_1^{k_1} \cdots p_r^{k_r}, \tag{1.17}$$

for any sequence of non-negative integers $k_1, \ldots, k_r$ such that $k_1 + \cdots + k_r = n$. The intepretation of the multinomial distribution is just a generalization of the experiment with three marbles. Consider a random experiment with r possible outcomes $s_1, s_2, \ldots, s_r$ and for each integer i, $0 \leq i \leq r$, let p_i be the probability of outcome s_i. Consider n independent repetitions of the experiment, and let Y_1 be the number of trials that result in outcome s_1, Y_2 be the number of trials that result in outcome s_2, etc. Then $(Y_1, \ldots, Y_r)$ has the multinomial distribution with parameters $(n, r, p_1, \ldots, p_r)$.

1.2.3 Populations and Sampling.

An important application of elementary probability is to model sampling from a population for the purpose of statistical inference. We have already introduced this type of application in examples 1 and 2. We will explain the basic terminology and context for studying a population through random sampling. Although you may not have encountered this in an elementary probability course, the probability we actually use will be familiar and simple. The important thing is to understand the set-up, as it is basic to the statistical analysis we shall introduce in the course.

The first thing to understand is the general notion of a population. Because this is a course in applications to biology, we will be modelling questions about real populations in the sense that we normally think of them—a collection of living organisms. The statistical notion of a population is more abstract. It is a collection of objects—people, or marbles, or DNA sequences, or pea plants—with measurable characteristics. Illustrative examples to keep in mind are our box of different colored marbles or the students at Rutgers labelled by weight and height. The latter example might arise in a study of the distribution of heights and weights in young adults. In this example, since we are only interested in heights and weights, we could think of the population abstractly as a collection of height/weight pairs, rather than as a population of actual students.

Consider now a population, each labelled by a single, real-valued measurement (such as height). Assume the measurement can takes a discrete value in a finite set $S = \{s_1, s_2, \ldots, s_r\}$—for example, a set of heights measured to the nearest inch. We shall conflate each individual with its measurement and talk about a population of numbers, each one belonging to the set S. The population average is the sum of all the measurements divided by the size of the population. To express this, let N_i be the number of s_i's in the population, and let N be population size. Then, since the N_i measurements of s_i contribute $s_i N_i$ to the total sum, the population average is

$$m = \frac{1}{N} \sum_{i=1}^{r} s_i N_i, \tag{1.18}$$

The population variance is the average square distance from the mean m. Similar reasoning implies it is given by

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{r} (s_i - m)^2 N_i. \tag{1.19}$$

Now consider sampling this population random. This means that we draw an individual from the population at random, *so that each individual has the same probability, namely,* $1/N$*, to be drawn.* Let the measurement labelling the individual drawn be denoted X. Clearly, X is a random variable with values in S and probability mass function,

$$p_X(s_i) = \frac{N_i}{N}. \tag{1.20}$$

By comparing the definitions of the expectation and variance of X to formulas (1.18) and (1.19), we see that

$$E[X] = m \qquad \text{and} \qquad \text{Var}(X) = \sigma^2. \tag{1.21}$$

This may appear all very obvious, but the set-up is fundamental to statistical estimation and testing. The central problem of statistics is to make inferences about and estimate the structure of an unkown population. We could of course examine every member of the population and record it, but this is typically impossible. Instead, we take a **random sample**, with replacement, of the population. That is we sample the population n times, each sample being random in the sense just described, and each being independent of the other. The sample is a sequence of independent, identically distributed random variables $X_1, \ldots, X_n$, each with the distribution specified by (1.20). Then we use the random sample to estimate the population. The simplest, natural estimates are the **sample mean**, defined by

$$\hat{m}_n \triangleq \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{1.22}$$

and the **sample variance**, defined by

$$\hat{\sigma}^2 \triangleq \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{m}_n)^2. \tag{1.23}$$

(The reason for dividing by $n-1$ rather than n will be taken up in the exercises.)

It is clear that the estimates should improve as n increases. This improvement and bounds on the accuracy can be obtained using Chebyshev's inequality (see the Appendix, II.10) Recall that Chebyshev's inequality states that for a random variable X

$$\mathbb{P}\left(\mid X - m_X \mid \geq a\right) \leq \frac{\text{Var}(X)}{a^2}. \tag{1.24}$$

We apply this to the sample mean. Note first that, using the linearity propery of expectation—see (1.13) and Appendix, II.5—(1.21), and $E[X_i] = m$ for every i,

$$E[\hat{m}_n] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} m = m; \tag{1.25}$$

the mean of the sample mean equals the mean. A second calculation, using (1.14), II.7 in the Appendix, and (1.21), yields

$$\begin{aligned} \mathrm{Var}(\hat{m}_n) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{\sigma^2}{n} \end{aligned} \tag{1.26}$$

Now apply the Chebyshev inequality using these results. We get

$$\mathbb{P}\left(|\hat{m}_n - m| \geq a\right) \leq \frac{\sigma^2}{na^2} \tag{1.27}$$

Since, for any positive a, this quantity tends to zero as n tends to infinity, the accuracy of $\hat{m}_n$ as an estimate of m increases.

Example 3. Consider a large box of red and white marbles. Let p be the probability that a red marble results from a random draw. Suppose we take a random sample of 200 draws, with replacement. Let $\hat{p}$ be the fraction of red marbles in our sample. Find a bound on $\mathbb{P}(|\hat{p} - p| > 0.1)$.

To put this problem in the set-up of the previous discussion, label each read marble by 1 and each white marble by 0. Then, p is the population mean, which we were calling m, and the sample mean $\hat{m}_{200}$ is just $\hat{p}$. Since a random draw X, resulting in 1 for red and zero for white, is a Bernoulli r.v., the population variance is $\sigma^2 = p(1-p)$. Presumably, we are in the situation where p is unknown so we really don't know σ^2. But since p is a number in $[0, 1]$, $\sigma^2 \leq \max\{p(1-p)\,;\, 0 \leq p \leq 1\} = 0.25$. Thus, from the bound (1.27),

$$\mathbb{P}\left(|\hat{p} - p| \geq 0.1\right) \leq \frac{.25}{200(0.1)^2} = 0.125.$$

(More refined techniques give much better estimates in this case.) ◇

The terminology, *random sample*, and the definition of sample mean and variance, are applied more broadly than to sampling from finite, if large, populations. Let p_Y denote a probability mass function on a set S of real numbers and let $m_Y \left(= \sum_{s\in S} s p_Y(s)\right)$ denote its mean, and variance $\sigma^2 \left(= \sum_{s\in S}(s - m_Y)^2 p_Y(s)\right)$. A **random sample from** p_Y is a sequence

$X_1, X_2, \ldots, X_n$ of independent random variables, each having p_Y as its probability mass function. The calculation leading to (1.27) for the sample mean goes through without change in this generalized setting. It is important enough to record it here as a theorem.

Theorem 1 *If $X_1, \ldots, X_n$ is a random sample from p_Y,*

$$\mathbb{P}\left(|\hat{m}_n - m_Y| \geq a\right) \leq \frac{\sigma_Y^2}{na^2} \tag{1.28}$$

Before leaving this subject, we briefly mention sampling without replacement. In sampling with replacement, the sampled individual is returned to the population before the next sampling. Thus there is a chance that an individual will be selected more than once in a random sample. One may also sample without replacement, so that once an individual is sampled it is removed from the population, if not physically, at least in the sense that it is not a candidate in further sampling. Computing ptobabilities for random samples without replacement is more complicated because the samples are no longer independent. What happens in the first sample changes the sampling population and so changes the probabilities of future random draws. For example, consider a box with two white and three red marbles. The chance of drawing white in the second sample is certainly different in the cases that a red is drawn first or a white is drawn first. However, if the population is large and the size of the sample is small relative to the population size, then each successive sample has only a small effect on the probabilities of future draws. In this case, it is valid to approximate sampling without replacement by sampling with replacement.

1.2.4 The Law of Large Numbers

The Law of Large Numbers is a basic result that goes to the heart of the frequency interpretation of probability. It concerns what happens to the sample mean of a random sample from a probability distribution as the number n of samples increases. We have already quantified how the sample mean becomes a more and more accurate estimate of the true mean in Theorem 1. By applying limits as $n \to \infty$ in this result and observing that the right hand side has limit 0 whenever a is positive, we obtain the following.

Theorem 2 *(***Weak Law of Large Numbers***) If $X_1, \ldots, X_n$ is a random sample from p_Y with a finite variance,*

$$\lim_{n\to\infty} \mathbb{P}\left(|\hat{m}_n - m_Y| \geq a\right) = 0 \quad \textit{for every positive } a. \tag{1.29}$$

This is called the Weak Law because it is a statement about limits of probabilities. But the following theorem about probabilities of limits is also true

Theorem 3 *(***Strong Law of Large Numbers***) Let $X_1, X_2, \ldots$ be pair-wise independent random variables, each with the sampe probability mass function p_Y. Suppose m_Y is the mean of p_Y and is assumed finite. Then*

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{1}{n}\sum_{1}^{n} X_i = m_Y\right) = 1. \tag{1.30}$$

This is a rather general statement of the strong law. Its proof is quite advanced and will not be given here.

Example: Frequency interpretation of probability. Consider a random experiment. Let A be an event in the space of outcomes of the experiment. Suppose, we run the experiment over and over. Define the frequency with which A occurs in the first n trials by

$$\hat{p}_n \triangleq \frac{\text{number of occurences of } A \text{ in first } n \text{ trials}}{n}.$$

This is the same empirical probability of A, estimated from the outcome of the first n trials. Theorem 3 says that

$$\lim_{n\to\infty} \hat{p}_n = \mathbb{P}(A) \quad \text{with probability 1.}$$

In words, the probability of A is the long run frequency with which A occurs in repeated independent trials. To see that this is true from Theorem 3, define the random variables

$$X_i = \begin{cases} 1, & \text{if } A \text{ occurs in trial } i; \\ 0, & \text{otherwise.} \end{cases}$$

Clearly, each X_i is a Bernoulli random variable with parameter $p = \mathbb{P}(X_i = 1) = \mathbb{P}(A)$. So they all have the same probabilty mass function and the mean, or expected value m is the expected value of this distribution, which

we know from (1.10) is $m = p = \mathbb{P}(A)$. By the assumption of independent trials, these random variables are independent. Also, $\sum_1^n X_i$ is equal to the number of times that A occurs in the first n trials. Hence,

$$\hat{p}_n = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

and so Theorem 3 implies

$$\lim_{n\to\infty} \hat{p}_n = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = m = \mathbb{P}(A).$$