Some Mathematics for Bioinformatics

Biological systems are, as a rule, large and complicated. This makes it unlikely that deterministic equations which describe these systems in detail will prove useful, because the complexity of these systems will be reflected in a similar complexity in the equations that describe them. Furthermore, since obtaining experimental measurements of the values of the many variables that would appear in such equations is difficult or impossible, we need to simplify our description somewhat if we intend to apply quantitative methods to biological systems. One way that we can accomplish this is by moving from an explicit to an implicit treatment of a great many of the variables that complicate these expressions. In doing so, we choose to model the collective influence of these "ignored variables" on the system by resorting to a probabilistic treatment of these systems. This turns out to be an important and useful simplification, and therefore a brief discussion of probability theory may prove useful to your future work in biology. In addition, it will provide a nice starting point for our later discussion of statistics.

We are used to thinking about probabilities as being a generalization of the notion of the frequency of occurrence of an event, where the probability that the event under consideration occurs is defined as the ratio of the number of favorable outcomes of an experiment to the total number of trials in the limit of an infinite number of trials.

$$P[A] = \lim_{N \to \infty} \frac{N_A}{N_{total}} \tag{1}$$

It turns out that this definition is unacceptable from a mathematical point of view, but it provides a fine guide to your intuition, and therefore you should not abandon it entirely. Additionally, if this definition is modified slightly, it proves to be very useful in the computation of probabilities of certain types of events (when there are only a finite number of equiprobable outcomes in the experiment), as we will discuss shortly.

Rather than the above definition, which is mathematically unacceptable because it relies explicitly on empirical observation, we will begin our discussion of probabilities using some notions from set theory. Though it sounds intimidating, set theory is nothing but a more formal version of logic, and you have already used notions from it innumerable times. Just to complicate things somewhat in the beginning, a set, technically speaking, can not be defined because any definition of a set must make use of the notion of a set and is therefore circular. However, technicalities aside, you can think of a set as a collection of objects, which are called elements.[1] As an example, the set of

[1] You also should be aware that a set may consist of just a single element. Such a set is called a *singleton.*

all people in Prof. Gerstein's Bioinformatics class is equivalent to the set composed of a list of your names. In addition to being an example of a set, the above also provides a nice illustration of an important notion from set theory, which says that a set may be specified by either providing a condition that all elements of the set satisfy or by enumerating every element of the set. For example, the set of all integers greater than 2 and less than 10 may be written as follows

$$S = \{x \mid 2 < x < 10 \text{ and } x \in N\} = \{3,4,5,6,7,8,9\} \quad (2)$$

where the bar should be read "such that" and the "fork" indicates that the quantity to the left are the elements of the set to the right. In addition, the set N is the set of all natural numbers (positive integers). Therefore, the above may be translated "the set S is equal to the elements x such that every element is greater than two and less than 10 and every element is an integer". A bit formal, but necessarily so-set theory is the distillation of logic and is therefore quite formal. In addition, there are two sets that are so important that they deserve special mention. They are the universal set Ω, which is the set that contains every element under consideration and the null set $\varnothing$, which contains no elements (and consequently is sometimes called the empty set).

Having defined sets, we can perform some basic operations with them. Given two sets A and B, we define three operations called intersection, union and complement. Introducing some more notation,

Definition of the Intersection of Sets

Given the sets A and B, we represent the intersection of the sets A and B as

$$A \cap B \equiv \{x \mid x \in A \text{ and } x \in B\}$$

Definition of the Union of Sets

Given two sets A and B, we represent the union of the sets A and B as

$$A \cup B \equiv \{x \mid x \in A \text{ or } x \in B\}$$

Definition of the Complement of a Sets

Given a set A and the universal set Ω, we represent the complement of A as

$$A^c \equiv \{x \mid x \notin A \text{ and } x \in \Omega\}$$

Extrapolating from this, the elements of a set are therefore also sets (with at least one element, but possibly more). Although such set theoretic musing is crucial for the formal development of modern probability theory, it is not as relevant to your work here.

A few words are in order about the above. The intersection of two sets produces a third set whose elements are only those elements that are common to both sets A and B. This is most easily understood using a Venn diagram, as shown below

Next, the union of two sets produces a third set whose elements are those that were contained in either set A or set B, or both, as below.

Lastly, the complement of a set A is the set whose elements are contained in the universal set Ω but not the set A, as below.

Having been introduced to the basic set operations, you should note that we can combine these operations and extend them to encompass a number of different sets (rather than just two). Therefore, we can define the intersection and complement of an arbitrarily large number of sets (even an infinite number of sets, though we will not discuss this case), and we can also combine these operations in a variety of ways. Some possibilities are illustrated below using Venn diagrams.

Finally, before moving on to probability theory proper, you should also note that we can use the above listed operations to provide a rigorous definition of a subset, which you

already know is just a set whose every element is also an element of another set. In terms of our new set theory notation,

Definition of a Subset

The set A is a subset of the set B if every element of A is also an element of B and is represented

$$A \subseteq B \; iff \; \exists \{x | x \in A \Rightarrow x \in B\}$$

The set A is called a proper subset of B if every element of A is also an element of B and the two sets are not equal. This is represented

$$A \subset B \; iff \; \exists \{x | x \in A \Rightarrow x \in B \; and \; A \neq B\}$$

The set A and B are defined as equal if they are subsets of each other,

$$A = B \; iff \; A \subseteq B and \; B \subseteq A$$

Where the iff is to be read "if and only if", the backwards "E" should be read "there exists" and the arrow should be read "implies that".

So what does all this set theory have to do with calculating probabilities anyway? The idea is that we can (tentatively) replace the aforementioned interpretation of probability, which requires actually doing an experiment a large number of times and counting up the number of times some outcome occurs, with a more sophisticated (and less labor-intensive) method. In this method, we count all of the outcomes that could occur in the experiment (if there are a finite number of outcomes) and then count all of the outcomes that are included in the event whose probability we are interested in calculating. Then, we can just divide the number of outcomes included in our "favorable event" set by the total number of outcomes possible. In other words,

(3) $$P[A]=\frac{\{A\}}{\{\Omega\}}$$

where {A} represents the number of sample points in the event set and {Ω}represents the number of points in the universal set.

Before we apply this however, a few more words about jargon are in order. It turns out that, due to historical accident, probability theory has a somewhat different terminology than does set theory. For example, in probability theory the universal set is called the *sample space*, and its elements are called *sample points*. Therefore, the sample points represent all of the possible outcomes of an experiment. Additionally, you should make special note of the fact that only one of these outcomes can occur per trial.

Therefore the sample points represent mutually exclusive outcomes. Using the notation of set theory, mutually exclusive sets are those sets which share no elements, i.e. A and B are mutually exclusive iff $A \cap B = \varnothing$. Such sets are called *disjoint* (their intersection is the null set). For example, in an experiment where we flip a coin once, the possible outcomes are heads and tails (only one of which may occur per flip). Therefore, the sample space is given by the set in which each element is a mutually exclusive outcome and every outcome of the experiment is represented by a sample point. Occasionally, you will find this last condition referred to as the exhaustive property of the sample space.

Now, back to our problem of calculating probabilities. As a slight increase in complexity from the last example, lets imagine an experiment where we flip two coins once. Then, the set of all possible outcomes (the sample space) is

$$\Omega = \{HH, HT, TH, TT\}$$

To illustrate the utility of the set theoretic approach to computing the probabilities of events (when the state space is finite), consider the probability of the event that you get at least one head in these two flips. Therefore, the probability of this event is

$$P[A] = \frac{\{A\}}{\{\Omega\}}$$

where the notation is the same as in (3).

The subset representing the event we are interested in is

$$A = \{HH, HT, TH\}$$

Notice that there are three points in the event set, and four in the sample space. Therefore, assuming all sample points are equiprobable[2] , the probability of getting at least one head in two flips is 3/4=0.75.

While this method of computing probabilities is very useful, it has two major limitations. Namely, it is limited to experiments with a finite sample space and whose sample points are all equiprobable. Unfortunately, it is frequently the case that we are interested in calculating probabilities of events that can not be represented in sample spaces satisfying both of these criteria.[3] However, we can extend the definition of probability much further by defining probability as a function which is defined on all subsets of the sample space that obey certain rules.[4] Although the details of this

---

[2] This assumption of the equiprobability of sample points is the major precept of the so-called classical definition of probability, which is the one we are now discussing. As you will shortly see, this definition of probability has been superseded by a more powerful and abstract definition.

[3] Consider the case of flipping an unfair coin, e.g. probability of heads=3/4, probability of tails=1/4. Clearly, even though the sample space is finite, the sample points are not equiprobable.

[4] In particular, these subsets must contain, in aggregate, every sample point in the sample space and must be closed with respect to finite union and complement. This collection of subsets is sometimes called a Boolean algebra.

definition of probability are quite involved, we can boil much of it down to a final definition of probability , as follows:

Definition and Axioms of Probability

Given a sample space in which each sample point has an associated probability (measure), the probability of any event is the sum of the probabilities of the sample points that are elements of the event set, or

$$P[A] = \sum_{x \in A} P[x]$$

Axiom I.

$$P[A] \geq 0 \; \forall \{A\}$$

where the upside down "A" should be read "for all"

Axiom II.

$$P[\Omega] = 1$$

Axiom III.

$$P\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} P[A_i] \quad \forall \quad A_l \cap A_k = \varnothing \; if \; l \neq k$$

These axioms are essential to the further discussion of probability and you should therefore be very familiar with them. Briefly, Axiom I. says that probabilities are non-negative. This is rather intuitive, since an event may happen (positive probability), or never happen (zero probability), but can not happen "less than never". Therefore, probabilities are necessarily non-negative for the same reasons that distances, areas and volumes are non-negative (you can never have less than no acres of land or less than no liters of water, etc.). This is not coincidental-probabilities are measures in the mathematical sense, just the same as distances, areas or volumes. Moving on, Axiom II. states that the probability of the state space is unity. All this means is that something must happen in your experiment, and therefore some point in the sample space (which contains all possible outcomes) must be realized. Lastly, and perhaps most importantly for the calculation of probabilities, Axiom III. says that the probability of a union of disjoint (mutually exclusive) events is equal to the sum of the probabilities of the events.

This property is sometimes called “countable additivity”, and is the starting point of all calculations of probabilities using set theory.

In closing these brief introductory notes on probability theory, we consider the notions of independence of events and conditional probabilities, which are essential to our further discussion of statistics. Intuitively, independent events are those events where the occurrence of one event does not alter the probability of subsequent occurrence of the other. For a more rigorous definition, we define independence as requiring that

$$P\left[\bigcap_{i=1}^{n} A_i\right] = \prod_{i=1}^{n} P[A_i] \text{ for independent events } A_i$$

where the capital pi is mathematical shorthand for the product. You have probably already been familiarized with this concept and computation rule, but it won’t hurt you to see it again. You should also verify that, in the event of equiprobable outcomes and a finite state space, the result of multiplying probabilities of independent events is identical to that obtained by the “ratio of sample points” method discussed earlier.

Having addressed independent events, we will now turn our attention to computing the probabilities of events which are not independent. For example, given that I roll two dice and the sum of the two faces is 7, what is the probability that I rolled a 6 and a 1 (in that order)? If I did not know the sum of the two faces, then the probability is just (1/6)x(1/6)=1/36 (for two fair dice), by independence. But I do know the sum, and therefore the outcomes of the two die are not independent. This is true because the sum must equal 7, which constrains the admissible outcomes of the experiment. However, we can calculate this probability using a familiar method if we look at the problem somewhat differently. Since not all of the original sample space is relevant to this case (only those outcomes which sum to 7 are), we could construct a new sample space, which is just a subset of the old sample space. This subset would consist of all outcomes of the two dice which sum to 7 (there are six of them, which you should verify), and then every sample point in our new space would be relevant. Then, with this subset as our new sample space, we see that the outcome 6,1 represents 1 of the 6 possible outcomes, and therefore the probability of getting this outcome given the foreknowledge of a sum equal to 7 is 1/6.

In more general terms, the probability of one event occurring, conditional on the outcome of some other event, is called a *conditional probability*. By way of notation, the probability of A given B is represented $P[A|B]$. Furthermore, we can visualize what a conditional probability represents by using set theoretic ideas to generalize the method we used above. The set representing the known outcome (B) represents the sample space. The event that A occurs given that B occurs is equivalent to the intersection of sets A and

B, i.e. $A \cap B$. Therefore, for a finite sample space with equiprobable outcomes, $P[A|B]$=(the number of elements in $A \cap B$/the number of elements in B).

However, this method, as used above to calculate a conditional probability suffers from the same limitations that we encountered before with probabilities defined on finite sample spaces and with equiprobable outcomes. We can circumvent these problems in a way exactly analogous to our previous solution; we will work with the probabilities of the subsets explicitly, rather than limiting ourselves to just considering the ratio of numbers of sample points in these sets. Therefore,

A Rule for Calculating Conditional Probabilities

For two event subsets A and B, the conditional probability $P[A|B]$ is given by

$$P[A|B] = \frac{P[A \cap B]}{P[B]}$$

Though we will end our discussion of conditional probabilities here, they occupy a central position in the machinery of probability theory and being able to manipulate them is essential to further study in the field.

After this somewhat rushed introduction to probability, you are now equipped to consider what may be the most important thought-experiment in the history of probability theory. Statements that grandiose always sound a bit intimidating, but fear not-one of the most charming aspects of this experiment is its simplicity. All that it involves is considering the probability that we get k heads in n flips of a coin.[5] However, for generality, we will consider a coin that may be loaded (i.e. unfair). By now you should immediately recognize this means that we can not use our "ratio of sample points" method because all outcomes in the state space are no longer equiprobable. However, outcomes of coin-flipping experiments are an excellent examples of independent trials, and therefore we can just multiply probabilities. Before progressing, lets take some time to formulate this problem in a bit more mathematically precise way. The experiment has two outcomes per flip[6], with the probability of heads=p and the probability of tails=q=(1-p). Furthermore, we will be flipping our coin n times, so each sample point is a list of the n outcomes of these flips. Therefore, there are 2^n sample points in the sample space (not all of which are equiprobable if $p \neq q$). Since the outcomes of each flip are independent, we can just multiply the probabilities for these two outcomes, and get

$$P[X = k] = p^k q^{n-k} = p^k (1-p)^{n-k} \qquad (4)$$

[5] Where n and k are arbitrary positive integers with $k \leq n$.

[6] Whether coin flipping or not, any experiment consisting of repeated, independent trials which have only two outcomes per trial and where the probabilities of these two outcomes are constant (over the course of the experiment) is called a Bernoulli trial.

The bad news is that this probability is dramatically less than the true probability. What have we done wrong? Recall that the sample points are lists of the outcomes of the n flips. Since order matters in a list, the results (hhhtth) and (hhthth) would represent two distinct outcomes (and therefore sample points), even though they have the same number of heads and tails. Therefore, the reason our calculation of the probability above is wrong is that we have not taken into account the fact that there are many different sample points included in the subset corresponding to obtaining k heads in n flips. Therefore, we need to multiply the probability above by the number of ways there are to get k heads in n flips (which is equivalent to the number of sample points in our event set).

Now we must confront a difficulty. We could count these outcomes by writing out every sample point (sequence of outcomes) for the flips, and then count those that have k heads, but this gets practically impossible for even a modest number of flips (remember the number of points in the sample space is 2^n). However, we can proceed by making use of some results from a field of mathematics called combinatorics. Looking at the problem somewhat differently, we are asking "how many different arrangements (mathematicians call them *permutations*) of k objects of one kind (heads) and n-k objects of another kind (tails) can we achieve"? The important point here is that we must consider each of the n flips as distinct, rather than just recording whether we got a head or a tail. To make this explicit, you might imagine that we have numbered every flip in addition to recording the outcome. This means that we have n choices for the first position, since any one of the n results can be placed in that position. However, for the next choice, we have only (n-1) possibilities, and for the subsequent choice (n-2), etc., since every choice depletes our "pool" of subsequent choices by one. Therefore there are n!=(n)(n-1)(n-2)...(1) permutations of n distinguishable objects.[7] However, consider the two permutations ($h_1t_2t_3h_4h_5$) and ($h_4t_3t_2h_1h_5$), where the subscript numbers indicate the numbered outcomes discussed above. From a permutation standpoint these are distinct outcomes, but for our purposes they are identical, since they both represent the sequence (htthh). Therefore, we have overcounted the number of distinct sequences because we have considered every outcome (every head and every tail) as distinguishable. Since we are only concerned with how many sequences of k heads and n-k tails there are, without regard to which head or which tail we are talking about, we need to divide n! by the number of ways that we can switch around all of the numbered heads and tails and still preserve the sequence. We now know that for k heads, there are k! permutations, and likewise (n-k)! permutations for the tails. Therefore, there are

[7] Products of the form (n)(n-1)(n-2)(n-3)...(1) are called *factorials*.

(5) $$\binom{n}{k} = \frac{n!}{(k)!(n-k)!}$$

different sequences of k heads and (n-k) tails. In combinatorial parlance, these distinct sequences with which we are concerned are called *combinations*, to distinguish them from the aforementioned permutations. This means that the true probability is

(6) $$P[X = k] = \binom{n}{k} p^k q^{n-k} = \binom{n}{k} p^k (1-p)^{n-k}$$

The expression above is a venerated distribution in discrete probability. It is called the *binomial distribution*, and it crops up so frequently in such diverse applications that we need no further justification for emphasizing it. However, in addition to being important in its own regard, the binomial distribution also marks a critical juncture in the development of probability theory. To see why, consider the presence of the factorial terms in the binomial coefficient. They are problematic since these terms grow rapidly as n increases, making calculation of the binomial coefficient challenging (especially in a pre-computer era) for n greater than 10 or so. Therefore, rather than explicitly evaluating this expression if n is large, we would like to find some function which closely approximates the binomial distribution when n is large. In addition to closely approximating the binomial distribution, this function may be continuous (for convenience) and should be valid within a wide range of values for p (the probability of "success" in our Bernoulli trial). Unfortunately, there is no single function which satisfies the last two of these criteria, but we can find a set of two functions (one continuous and one discrete) that satisfy all of these criteria. These functions are the *normal (Gaussian)* and *Poisson* distributions (respectively), and we will shortly turn our attention to both.

However, before moving on to discuss these distributions, we should take some time to briefly discuss a new idea that was introduced with the binomial distribution. Recall that we were concerned with calculating the probability of getting k heads in n flips of a coin. Though there was a subset of sample points which satisfied this criterion, we needed to recognize how the variable "get k heads in n trials" depended on these sample points. Similarly, we might have discussed the probability that I make $5 if I make $0.50 on each head and the coin is tossed n times. While this quantity depends on the outcomes of the flips (sample points), it is not one of them. Therefore, we have introduced the idea of a quantity which may be regarded as a function of the sample points, where the sample points act as the independent variable. Such a quantity is called a *random variable*, and it is essential that you be familiar with this concept before we move on to discussing various probability distributions. To gain some practice with the

concept, imagine some other random variables that we could use in conjunction with the binomial distribution. Hopefully, you will appreciate that there are an infinite number of them, and therefore you should expect that most applications of probability theory to real life involve dealing directly with random variables rather than sample points.

Having introduced the idea of approximating the binomial distribution with two distributions, each of which is applicable in a different regime of the value of p, lets consider the case where p is small ($p \leq 0.1$). First, let us perform the substitution $\lambda=np$.[8] The binomial distribution then becomes,

$$P[X=k]=\binom{n}{k}\left(\frac{\lambda}{n}\right)^{k}\left(1-\frac{\lambda}{n}\right)^{n-k} \tag{7}$$

Now, consider the case where n grows to infinity and p shrinks to zero. Hopefully you appreciate the utility of the substitution that we made above, since we can force n to grow and p to shrink such that $\lambda=np$ remains constant. This is nice since nothing in the above expression will "blow up" for large n and/or small p. In this limit[9] we get,

$$f(k)=\frac{\lambda^{k}e^{-\lambda}}{k!} \quad \text{for k=0,1,2,3,...} \tag{8}$$

This expression is the Poisson distribution, and is useful in the situations where the probability of an occurrence is small and the number of "trials" (n) is large. For example, we might consider the probability of k adverse reactions to a test drug in a given sample of the population or the probability of registering k complaints about a particular product in a 1-hour period or the probability of finding k point mutations in a given stretch of nucleotides. Though the Poisson distribution is essential to application and you will doubtless see it again, we will leave it now to discuss the other binomial-approximating continuous distribution.

If the probability of success in a Bernoulli trial is not sufficiently small to justify the assumptions that went into the derivation of the Poisson distribution, then we may derive an alternative approximating distribution. It can be shown that[10] this distribution is continuous and given by

$$f(X)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{(X-\mu)}{\sigma}\right)^{2}} \quad \text{for } -\infty < X < \infty \tag{9}$$

---

[8] The value of λ, called the *parameter* of a Poisson distribution, is required before you can calculate probabilities using this distribution. Often times, λ must be empirically determined.

[9] I will omit the detailed derivation, but for the adventurous among you, keep in mind the dual limits and note that $(1-\lambda/n)^{n}$ converges to $e^{-\lambda}$ for large n.

[10] It can be shown, but isn't, because the detailed proof of the normal distribution as a limit of the binomial distribution is rather involved. If you are interested in further reading on this derivation, you should start with the Local and DeMoivre-Laplace limit theorems.

This is the normal (or Gaussian) distribution. You are already familiar with this distribution from course grading (as well as much else), and it looms over probability theory as the single most important continuous distribution in applications. It is so significant that it can be proven that every distribution approaches a normal distribution under suitable limits. [11]

As where the Poisson distribution was characterized by a single parameter, λ, the normal distribution is characterized by two quantities; the *mean* μ and the *variance* σ^2. The mean is just a generalization of the notion of an average, whereby we weight every value that the random variable (r.v.) can attain by the probability that this particular value is realized.[12] In other words,

$$\mu = \sum_{i=1}^{n} x_i P[X = x_i] \tag{10}$$

where x_i are all of the values that the r.v. can attain and $P[X = x_i]$ is the probability that the r.v. achieves a value x_i. In addition, you should verify that in the case that all values of the are equiprobable, the mean reduces to the arithmetic average. Therefore, the mean is just a weighted average. Furthermore, the mean is often called a "measure of location", which makes sense given that the normal distribution is centered around the mean. However, for non-symmetric distributions, there are two other measures of location that become non-redundant with the mean. These alternative measures of location are the median and the mode. The median is the value of the r. v. where there are equal numbers of observations with values greater and less than the median. The mode is the most frequently observed single value of the r. v. These are of little relevance to the normal distribution, but are often of significant value when dealing with data for statistical purposes.

Whereas the mean provides information concerning the location of the distribution, the variance provides information concerning the "spread" of the distribution about the mean. In particular,

$$\sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 P[X = x_i] \tag{11}$$

You may be more familiar with the standard deviation, which is related to the variance as below.

$$\sigma = std.dev. = \sqrt{\sigma^2} \tag{12}$$

---

[11] This is a hand-waving version of the Central Limit Theorem, which says that the distribution of the sum of independent, identically distributed (i.i.d) random variables (r.v.) will approach a normal distribution in the limit of an infinite number of i.i.d r.vs.

[12] Particularly in physical applications, the mean of a r. v. is called an *expectation value*.

Thus, the standard deviation (std. dev.) is just the average difference between the mean and the observed values of the r.v. Therefore, if we are considering a set of observations that we believe are normally distributed, then the std. dev. (or the variance) provide a measure of the variability of observations.

Having introduced the normal distribution, we have yet to deal with a problem created by our shift from the discrete binomial distribution to the continuous normal distribution. Specifically, while we can speak of the probability that a r.v. equals a particular value with a discrete distribution over a finite sample space, we can not do so when dealing with a continuous distribution. This is because a continuous distribution has an infinite number of sample points, and therefore the probability that one value is realized is always zero (because you have one point in your event set and an infinite number in your sample space). Therefore, when dealing with continuous distributions, we can only speak of the probability that the r.v. takes on a value in a certain range. Therefore, taking the normal distribution as an example,

$$(13)\quad P[a \le X \le b] = \int_a^b f(X)\, dX = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}\left(\frac{(x-\mu)}{\sigma}\right)^2} dX = F(b) - F(a)$$

We see that to determine the probability that the r. v. falls within a range we must integrate these distributions. The functions F(a) and F(b) are called *cumulative distribution functions*, and are just integrated probability distributions[13]. The function f(X), which we have thus far called a probability distribution, is more properly called a *probability density*. This name is suggestive of an analogous shift from a discrete to a continuous treatment of mass in physics. For a small number of particles, we can treat matter as comprised of discrete mass points. For larger numbers of particles, we abandon this approach in favor of describing matter using a continuous function called a mass density. Similarly, we describe continuous probability distributions using probability density functions. To extend the aforementioned definitions of mean and variance to the continuous case,

$$(14)\qquad \mu = \int_{-\infty}^{\infty} X\, f(X) dX$$

$$\sigma^2 = \int_{-\infty}^{\infty} (\mu - X)^2\, f(X) dX$$

These expressions look very similar to expressions for the center of mass and moment of inertia for continuous mass distributions from physics. You now know that this not coincidental.

---

[13] Cumulative distribution functions are of the general form

We have covered enough probability theory at this point to begin our discussion of statistics. Probability and statistics are complementary approaches to the same problem of describing phenomena which are recalcitrant to deterministic treatment. Probability theory asks "given a model (probabilistic) of this phenomena, what can be said about the data that I might expect from experiments (or observations) performed on this system"? Statistics, on the other hand, asks "given data generated from experiments (or observations) on this system, what can I say about a model for this phenomenon"? Hopefully, you appreciate the close kinship of these two fields of study and are consequently full of forgiveness for the preceding 12 pages of introductory probability theory.

Statistics is a big field, and therefore we will be able to address only one small topic within the space of these notes. This is meant both as an apology and encouragement, since there is much more to statistics and you should be familiar with much of it for future work in biology. Apologies aside, the topic that will occupy our attention here is that of *hypothesis testing*. The name is rather self-explanatory, but the core idea is that we formulate some hypothesis about a population, take a sample of this population, and use the resulting characteristics of this sample to determine the probability that our hypothesis is correct. The hypothesis that we are testing is called the *null hypothesis*. To test this hypothesis, we take a sample of the population and decide on a *test statistic*. The test statistic quantifies the agreement between the hypothetical characteristic of the population and the observed characteristic of the sample. Lastly, you compute the probability that your test statistic would have at least the observed value if your hypothesis (the null hypothesis) is correct. The probability is often called a *p-value*, and if it is below a pre-determined cutoff (usually either 0.05 or 0.01), you decide to reject the null hypothesis. In rejecting the null hypothesis, you are saying that the difference between the hypothesized characteristic of the population and your sample characteristic is too large to be due to "random" fluctuation, and therefore something else that is not accounted for by your null hypothesis must be at work.

To make all this a bit more concrete, lets consider an example. Suppose you are interested in determining whether a biotech. company is being honest about the activity of an enzyme they are selling. You purchase some enzyme, remove n aliquots and assay each aliquot for activity. The company claims that their enzyme activity is μ units/ml. Therefore, your null hypothesis is that the enzyme's mean activity is μ units/ml, and is normally distributed about this mean. Therefore, your test statistic in this case is

(15) $$Z_{obs} = \frac{\left(\overline{X} - \mu\right)}{\left(\sigma/\sqrt{n}\right)}$$

Where $\overline{X}$ is the sample mean, and σ is the sample standard deviation.

Essentially, we can view the test statistic as a r.v. which we assume is normally distributed.[14] Therefore, given a computed test statistic, we can use the normal distribution to calculate the probability that the null hypothesis is true. We will do this by computing the probability that the test statistic assumes a value at least as large as that observed. In other words,

(16) $$\text{p-value} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z_{obs}} e^{-\frac{z^2}{2}} dz = F(z)$$

If the p-value is less than a given cutoff (0.05 or 0.01), we conclude that it is improbable that the test statistic would assume such a large value and therefore reject the null hypothesis. In the case of our example, we may find a p-value<0.01, and therefore reject the company's claim concerning the activity of their enzyme. We should note in passing that the details of computing the p-value can vary. There exist two distinct methods of computation; the one-tailed or two-tailed tests. They differ only insofar as the latter test explicitly capitalizes on the symmetry of the standard normal distribution, and consequently, the details of the implementation are slightly different. You need not be overly concerned about this.

In closing our discussion of hypothesis testing, you should note that a test statistic can be constructed to test a wide variety of hypotheses. For example, we could construct a test statistic to determine whether the observed proportion of defective products in a sample is significantly different that the quoted defective rate. Many other such applications exist, and consequently hypothesis testing is among the most commonly employed of statistical methods.

With all this talk about probability and statistics, you may be fearing that deterministic mathematics is totally irrelevant to bioinformatics. This is most certainly not the case. In fact, a mathematical object of great utility in this field is the vector. This is because vectors provide a uniquely powerful way to represent and manipulate large quantities of information. For example, the structure of a biomacromolecule may be

---

[14] i.e. with probability density function $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$. In passing, r.vs with probability density functions like this are said to the distributed according to the *standard normal* (mean=0, variance=1). By performing a straightforward substitution, all normally distributed r.vs can be transformed to a standard normal distribution. This substitution is frequently called the *z-transformation.*

thought of as a 3N-dimensional vector, where N is the number of atoms in the structure. In what follows, we will review some of the most important aspects of vector algebra.

As you recall from physics, vectors are objects that have both magnitude and direction, and are many times represented by arrows to indicate this fact. In doing so, we should see that we can also represent this vector by its components along the x,y and z axes Therefore, in general we have

$$\vec{V} = V_x\hat{i} + V_y\hat{j} + V_z\hat{k} = \left\|\vec{V}\right\|\cos\alpha + \left\|\vec{V}\right\|\cos\beta + \left\|\vec{V}\right\|\cos\gamma \tag{17}$$

Where $\left\|\vec{V}\right\|$ represents the length (also called magnitude or norm) of our vector, $\hat{i}, \hat{j}, \hat{k}$ are the unit vectors (vectors of length=1) that point in along the x, y and z axes respectively, and α, β, γ are the angles between the vector and the x, y, and z axes, respectively (as below).

The above illustration shows us that the x,y, and z components of the vector are equivalent to the projection of this vector onto these axes. Since these projections are just the magnitude of the vector multiplied by the cosine of the angle between the vector and the axes, we realize that if we know the length of the vector and the angle it makes to each axis, then we know in what direction to draw the vector. Because these angles tell us in what direction to draw the vector, the cosines of these angles are called the *direction cosines*. They arise frequently and you should keep in mind that the direction cosines are just another way of representing the components of the vector under consideration. Before we leave the topic of direction cosines, consider the expression for the magnitude of a vector, shown below

$$\left\|\vec{V}\right\| = \sqrt{\left(V_x^{\,2} + V_y^{\,2} + V_z^{\,2}\right)} \tag{18}$$

This implies

(19) $$1 = \cos^2\alpha + \cos^2\beta + \cos^2\gamma$$

which you should verify.

Just as we can (uniquely) specify a vector by listing its components, we can also operate on vectors via operating on their components. For example, you learned in physics that vectors can be added geometrically using the "head-to-tail" rule for vector addition. Now, if what we just said about operating on vector component-wise is true, it should provide us with the same answer for addition of vectors as does "head-to-tail" addition. We verify this geometrically below

Therefore $\vec{A} + \vec{B} = (A_x + B_x)\hat{i} + (A_y + B_y)\hat{j} + (A_z + B_z)\hat{k}$. We can see that as the number of vectors increases, it becomes much more difficult to draw them all out and do "head-to-tail" addition. In contrast, adding 10 numbers is not much harder than is adding 2. Therefore, the great advantage of component-wise addition of vectors is that it allows us to reduce the potentially tedious problem of adding arrows to the much easier one of adding numbers.

Having conquered vector addition, you may be tempted to apply similar methodology to vector multiplication. Unfortunately, multiplying vectors is trickier than adding them, since there are two ways of forming vector products. One way is very similar to our component-wise method of vector addition, and it produces a number (scalar). Therefore, this method of vector multiplication is called the scalar (or dot) product. The other way of multiplying vectors is very different from straightforward component-wise operation and produces another vector. Consequently, it is called the vector (or cross) product. In what follows, we will take some time to discuss each kind of vector product and its significance.

The dot product of two vectors is given by the following relation

$$(20) \qquad \vec{A} \bullet \vec{B} = \left(A_x B_x\right) + \left(A_y B_y\right) + \left(A_z B_z\right) = \left\|\vec{A}\right\| \left\|\vec{B}\right\| \cos\theta$$

Where θ is the angle between these two vectors. The first equality above is useful primarily when actually computing a dot product. The second equality is useful when manipulating vector relations (although it is of considerable computational utility as well). The cosine term in the above suggests a geometric interpretation of the dot product very similar to that given for the direction cosines. Namely, this equality tells us that the dot product of two vectors is the product of the vector magnitudes projected along the direction of one of the vectors. Furthermore, the cosine term allows us to conclude that parallel vectors produce maximal dot products and perpendicular vectors produce zero dot products. Perpendicular vectors are called orthogonal, and are of very general importance. You will see them again in linear algebra, but with a somewhat less geometric interpretation. As a closing comment on the many-faceted wonder of the dot product, you should note that, by virtue of the first equality, the dot product is commutative (order of multiplication does not matter). This is a direct consequence of the commutativity of scalar multiplication.

In contrast to the appealing simplicity of component-wise multiplication that characterizes the dot product, the cross product is a less straightforward matter. We define the cross product as

$$(21) \qquad \vec{C} = \vec{A} \times \vec{B} = \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ A_x & A_y & A_z \\ B_x & B_y & B_z \end{vmatrix} \therefore$$

$$\vec{C} = \left(A_y B_z - A_z B_y\right)\hat{i} + \left(A_z B_x - A_x B_z\right)\hat{j} + \left(A_x B_y - A_y B_x\right)\hat{k}$$

where the bars on either side of the matrix indicate that we are to find the determinant of the matrix inside. Although we are dealing with a topic that more properly belongs in a discussion of linear algebra, lets discuss how to compute the determinant of a matrix.

For most "by hand" computations, a determinant is most easily found by using the so-called "expansion by minors" approach. This is done as shown below

(22) $$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = a\begin{vmatrix} e & f \\ h & i \end{vmatrix} - b\begin{vmatrix} d & f \\ g & i \end{vmatrix} + c\begin{vmatrix} d & e \\ g & h \end{vmatrix}$$

This should seem less than helpful, since we have just broken down one determinant into three. However, finding the determinant of a 2x2 matrix is rather easy. We just multiply the entries lying on "left-to-right" diagonal and subtract the product of the entries lying on the "right-to-left" diagonal. In other words,

(23) $$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

Knowing this, we can introduce a convenient mnemonic device for the computing the determinant of a 3x3 matrix, which we can represent as below

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = a\begin{vmatrix} e & f \\ h & i \end{vmatrix}$$

(24) $$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = -b\begin{vmatrix} d & f \\ g & i \end{vmatrix}$$

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = c\begin{vmatrix} d & e \\ g & h \end{vmatrix}$$

Knowing this, you should expand the determinant of the matrix in (21) and verify that you do indeed get the vector indicated in the second equality in (21).

Now that you're comfortable with how to compute the cross product, a few geometrical considerations are in order. First, the cross product always produces a vector which is perpendicular (orthogonal) to both of the vectors "crossed" in the cross product, as shown below.

In addition, the definition of the cross product in (21) leads us to conclude that the cross product, unlike the dot product, in not commutative. In fact, if we reverse the order of the vectors in (21) and compute the result, we get a vector which points in the opposite direction as the original product. More succinctly,

$$\vec{A} \times \vec{B} = -\left(\vec{B} \times \vec{A}\right) \tag{25}$$

This property is called anticommutativity. The direction of the resulting vector can be easily remembered using the right-hand rule. The rule says that the product vector will point in the same direction as the thumb on your right hand if you align your palm with the first vector and curl your fingers toward the second vector in the cross product. You probably were introduced to this in physics. In concluding our discussion of the cross product, you should be aware of the following relation between the magnitude of the product vector and the magnitudes of the vectors being crossed. Namely,

$$\left\|\vec{C}\right\| = \left\|\vec{A}\right\|\left\|\vec{B}\right\| \sin\theta \tag{26}$$

You will not make as much use of this as you will (20), but (26) is still worth remembering.

Furthermore, just as we can multiply more than two scalars, we can also multiply more than two vectors. However, always keep in mind that the two ways of multiplying vectors (dot and cross product) that we have just discussed are only defined for vectors. So, while we can make the triple product $\vec{D} = \vec{A} \bullet \left(\vec{B} \times \vec{C}\right)$, we can *not* make the triple product $\vec{D} = \vec{A} \times \left(\vec{B} \bullet \vec{C}\right)$, since the term in parenthesis is a scalar, and the cross product is only defined between two vectors. Because this requirement removes the potential ambiguity in the triple product, parenthesis are usually omitted. The triple product you are most likely to encounter in future studies is the scalar triple product, given by

$$V = \vec{A} \bullet \vec{B} \times \vec{C} \tag{27}$$

The choice of V to represent this scalar triple product is suggestive. In fact, the scalar triple product is often used to calculate volumes of parallelepipeds. A biophysically relevant example would be the volume of the unit cell of a crystal, which is a very important parameter in crystallography. However, this is only one of many applications of vectors to biophysics, so please make sure that you are comfortable with them before you continue.

Having treated vectors via their components, it will come as no surprise to you that we can further "economize" our treatment of vectors by representing them as lists of components. For example, we could represent the vector $\vec{V}$ in any of the three ways shown below.

$$\vec{V} = V_x\hat{i} + V_y\hat{j} + V_z\hat{k} = \begin{bmatrix} V_x & V_y & V_z \end{bmatrix} = \begin{bmatrix} V_x \\ V_y \\ V_z \end{bmatrix} \tag{28}$$

The last two quantities are examples of a *row vector* and a *column vector*, respectively. This may seem like dry exposition on bookkeeping methods for vectors, but it allows us to achieve a great simplification in how we handle vectors, to which we turn our attention now.

Consider the vector shown below, which is represented in two coordinate systems, one rotated with respect to the other by ϕ. The problem is to provide expressions for the components of the vector in the new (rotated) coordinate system in terms of its old components.

You will need to use the sum formulas from trigonometry to do this problem (and might try it as an exercise). In doing so, you will find the following;

(29)
$$V_x^{'} = V_x \cos\phi + V_y \sin\phi$$
$$V_y^{'} = -V_x \sin\phi + V_y \cos\phi$$

As you have already come to appreciate, this problem is somewhat tedious if solved in the manner suggested above. The great attraction of representing vectors as row or column vectors is that we can operate on these vectors using matrices. However, to do so, we will first need to discuss the anatomy and mechanics of matrices and their manipulation.

A matrix is simply an array of numbers, which are called entries. The entries are organized amongst rows and columns. As you have probably already anticipated, columns run up-down and rows run side-to-side. Therefore, we can uniquely specify an entry in the matrix by providing its row and column index. A generic matrix is shown below

(30)
$$\langle A \rangle = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

The above matrix is an example of a 3x3 matrix, where, by convention, the first number refers to the number of rows and the second to the number of columns. Matrices may come in any size, 1x3 (3-dimensional row vector), 3x1 (3-dimensional column vector), 3x3, 4x2, 100x1734, etc.

We can add matrices entry by entry, as below;

(31)
$$\langle C \rangle = \langle A \rangle + \langle B \rangle = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} = \begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} & a_{31}+b_{13} \\ a_{21}+b_{21} & a_{22}+b_{22} & a_{23}+b_{23} \\ a_{31}+b_{31} & a_{32}+b_{32} & a_{33}+b_{33} \end{bmatrix}$$

Note that this is exactly analogous to component-wise addition of vectors.

Continuing this analogy with vector algebra, while we can add (and subtract) matrices easily, multiplication is somewhat trickier. In brief, matrix multiplication obeys the following rule

Matrix Multiplication

Given two matrices $\langle A \rangle$ (which is MxN)and $\langle B \rangle$ (which is NxP), their product can be found as follows:

The element c_{ij} is the given by $c_{ij} = \sum_{k=1}^{N} a_{ik} b_{kj}$. The product matrix $\langle C \rangle$ is MxP. Note that matrix multiplication is *not commutative* (in general).

All that this definition is saying is that the element c_{ij} is just the i^{th} row of $\langle A \rangle$ multiplied by the j^{th} column of $\langle B \rangle$ in a component-wise fashion and then summed. You might recognize that this is similar to the method of computing the dot product of two vectors. As it turns out, there is an excellent reason for this similarity because this row-by-column method of matrix multiplication is perfectly equivalent to the dot product. To verify this, choose two vectors and compute their dot product. Next, arrange the components of first vector as a row vector, and the second as a column vector. Now, perform matrix multiplication as defined above and verify that the results are identical. In linear algebra, this is called the *inner product*.

Having defined matrix multiplication, we can now re-inspect (29) and recognize that this is equivalent to the matrix equation

$$\begin{bmatrix} V_x' \\ V_y' \end{bmatrix} = \begin{bmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{bmatrix} \begin{bmatrix} V_x \\ V_y \end{bmatrix} \tag{32}$$

The type of matrix shown in (32) is among the most useful matrices you will ever encounter. It is called a *rotation matrix*, and given the example that produced it, you should immediately appreciate why. These matrices are used in every imaginable application, and being comfortable with them is critically important.

Now that you've seen a rotation matrix in two dimensions, we will generalize this result to three dimensions. The key to doing this is to imagine rotating a Cartesian (3-D) coordinate system about one of its three orthogonal axes, as below.

This is equivalent to rotating the xy plane through an angle ϕ while leaving z unchanged. The three dimensional rotation matrix for this operation is

$$\langle R_z \rangle = \begin{bmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{33}$$

Note that, in matrices of this sort (in a Cartesian basis, to be precise), the first column represents the x-axis, the second the y and the third the z-axis. Here, we are assuming that we are working in a column space.

Now, we could describe an arbitrary three dimensional rotation matrix by considering the consecutive rotations of the coordinate system around z, x, and the "new" z-axis (new in that its orientation has changed with respect to its original position). Such matrices are necessary to describe the orientations of rigid bodies in space relative to some external coordinate system. Therefore, in matrix notation,

(34)

$$\langle R\rangle = \langle R_A \rangle\langle R_B \rangle\langle R_C \rangle = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\varpi & \sin\varpi \\ 0 & -\sin\varpi & \cos\omega \end{bmatrix}\begin{bmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The angles through which we are rotating the coordinates are called the Euler angles. However, two points about the above should be hastily made. First, since matrix multiplication is not commutative (order matters), the order that you perform these rotations is very important. Nearly every field that uses Euler angles has a different convention for order of rotations, so we will not waste space here discussing them all. In addition, this topic is among the most non-intuitive you are likely to encounter with any regularity. Therefore, if you're interested, you should curl up with some classical mechanics textbook and devote a hour or so to serious study. Lastly, we should note that we have just scratched the surface of the elegant field of linear algebra, which finds no end of application in every field of science. It is mandatory that you master elementary linear algebra in order to fully enjoy the mass of biophysically and bioinformatically relevant mathematics.