Probabilities and Such

Thinking about probabilities is both easy and hard. It's easy because we are surrounded by chance events and we are used to thinking about them. On the other hand, it's hard because we are used to thinking about them intuitively and we have many preconceptions...There are many ways to think probabilistically and we will for the most part ignore the formal definitions. To set things up, we will discuss a canonical situation....the throw of a dice.

The first thing we have is the probabilistic **EXPERIMENT**, that is, we are going to take a six sided square and throw it in the air and look at it when it settles. Next, we have a **TRIAL**, that is, one execution of the experiment we've setup. This implies that the experiment is inherently repeatable. It is somewhat a point of philosophical debate as to whether probabilistic concepts are applicable when we have an experiment that is not repeatable (such as evolution). The next concept is the **EVENT**, that is, the chance outcome of the trial. An important point to consider is that an event can be anything (well almost), not just a number--in fact, most often it is not a number. For example, when we toss a dice and get the face with six dots turned up, this is the event--a six sided square object with six dots on the top, not the number 6. The number 6 is a label we attach to the event. An event can be collection of other events as well and this will be more carefully examined below. For the moment I will be use this word in a casual manner. Obviously, for a probabilistic experiment we will have a collection of all possible events. This is the **EVENT SPACE,** or **SAMPLE SPACE** (more commonly used) or sometimes called the **UNIVERSE**.

I just mentioned that the number "6" is a label for the event "six dots on the top". This kind of numerical labeling is often used because we know quite well how to handle numbers but not really how to handle actual events (e.g., the sum HEADS+TAILS is not easily defined). We can label the entire sample space with numbers, in fact with real numbers. (This is called constructing a function from the event space to the "real line".) As you well know, the real line can be represented by a variable, say X. The variable X will take specific values on the real line. When this variable is a labeling scheme for the sample space we call this the **RANDOM VARIABLE**. For example, we might have a random variable X that takes the values {1, 2, 3, 4, 5, 6}. We could say that the random variable X represents the sample space of the experiment "dice toss" and the particular values of X represents each event (e.g., X = 3 might represent the event "three dots on top"). Since we see this as a labeling scheme, there is no reason that X should take the values {1...6}. It could have been X={10, 100, 150, 500, 550, (800-999)}. Also, there is no reason that the number X=1 should represent the event "one dot on top", it could mean "six dots on top". For the most part, much of the things we do in probabilistic modeling are done with random variables rather than the events themselves. We do have to constantly remind ourselves that the random variable is not the probabilistic event. It is just a marker for a probabilistic event.

So far, we haven't mentioned probability itself. This is a number attached to each event in the event space. Given the idea of experiments and the trials above, it is tempting to define the probability of an event as the frequency of the event after many repeated trials of the experiment. This definition works pretty well for the most part but there are difficulties, so we will take a formal definition. A **PROBABILITY** is a positive number we attach to each event in the sample space such that the sum of the numbers for all of the event space equals 1 (when we say "each event" or "sum of the numbers" I am being very imprecise, this will be made more precise below). This definition can also get us into all kinds of trouble and we will discuss them eventually, however, it works well for most purposes. But, we next discuss how to enumerate events.

1. Counting

Counting things, e.g., all the different ways of writing a four-letter word, is an important computation in probabilistic modeling. One of the reasons for this is that there is a particular kind of probability model called the equiprobable model. Under this model we declare that each elementary event in the sample space is equally probable. If there are N possible elementary events in the sample space, each event has a probability of 1/N. There are many ways to describe the sample space and one way is simple enumeration, e.g., the event space of a dice toss {"one dot", "two dots", "three dots", "four dots", "five dots", "six dots"}. If we are given such enumeration and we use the equiprobable model, we know that probability of "one dot" = 1/6. Often, the event space is described more concisely, e.g., {all possible four letter words}. This is why we need counting techniques.

(1) All possible four letter words and its ilk.

What want to do to construct all possible four-letter words is to take a label set, i.e., the alphabet, (but could be colors, numbers, names, taxa, DNA, whatever) and assign labels to four "slots", i.e., position for the letters (slots could be also urns, pigeon holes, marbles, balls, postboxes, sequence positions, whatever). Since our purpose is counting, we have to think about what we are counting (seems redundant, eh?). We note that in this example we are allowed to take many copies of the same label; that is, a word like AAAA is cool. Another thing is that we consider words like ABCD to be different from DCBA; that is, the order matters. So we can represent the problem as {(position 1) = letter,,,(position 4) = letter} or the number of ways of matching four label slots (position 1 to 4) to 26 labels (A..Z) allowing redundancy. (I know you knew the answer way before and this is deadly boring, but there is some use in carefully analyzing the structure of problems from the beginning.) The answer to the original problem is simple now, there are 26 different assignment at each position so we have 26*26*26*26=456976 ways and under the equiprobable model each four-letter word has a probability of 1/456976.

In general, we have k^n ways of assigning n labels to k (labeled) slots with redundancy.

So now, how many different sequences are possible for a 1kb of DNA? We have 1000 slots and four possible labels for each position -- $4^{1000} = 2^{2000} \sim 10^{200}$. To put this number

in perspective, there are supposed to be about 10^{80} elementary particles in the entire universe.

(2) Permutations

Simply, a permutation is all possible ways of ordering *n* items. For example, all possible permutation of ABC is {ABC, ACB, BAC, BCA, CAB, CBA}. We can also look at this problem as taking *n* labels and assigning it to exactly *n* slots without allowing for redundancy. In the ABC example, we first have three different labels {A, B, C} and three slots {1, 2, 3}. We can assign any of the three labels to, say slot 1. However, once we do that we only have two labels left for slot 2. Finally, once we decide slot 2, only one label is left for slot 3. So, as you might already know we have 3*2*1 = 6 ways of permuting ABC.

In general, we have $n(n-1)(n-2)...3\cdot 2\cdot 1 = n!$ ways of permuting (ordering) *n* things. (This is the meaning of the factorial (!) notation.)

Another way to think of permutations is as an assignment of *n* labels to *n* other labels. This is a very general view. Many problems in which we connect *n* things to *n* other things can be seen as a permutation..

How many different permutations are possible for a 1kb of sequence? This question immediately shows us the subtlety of counting. Let's start with a small example of a sequence of length 3. Suppose this sequence was ATC, then we have ACT, TAC, TCA, CAT, etc., for a total of 6 possible permutations. But, suppose we start with the sequence AAT, then we have the following possible shuffling of the positions: ATA, AAT, ATA, TAA, TAA. As you can see some of the possible permutations of the sequence are identical to each other. Is there a counting formula to deal with this situation? Yes! First note that if all the nucleotide in each position were different from each other we have 3! = 6 possible permutations. Now note that in the second case we have two nucleotides, the first and second A, that is identical to each other. That means that all the possible ways of switching the first A with the second A doesn't change the result. All the possible ways of switching the first A with the second A is 2!. So out of 3! possible permutations, there are 2! identical labelings and we have 3!/2! = 6/2 = 3 possible permutations.

Going back to the 1kb of sequence, if we have 250 A's, 250 C's, 250 G's, and 250 T's in the sequence, we would have $\frac{1000!}{250!250!250!250!}$ possible permutations. Now, computing numbers like 1000! is pretty hard. Fortunately, we have an approximation for these factorial computations. It is called the Stirling's approximation and it looks like:

$$n! \approx \sqrt{2\pi n}\cdot e^{-n} n^n$$

Using this approximation we have

$$\frac{1000!}{250!250!250!250!} \approx \frac{\sqrt{2\pi 1000}\cdot e^{-1000}1000^{1000}}{(\sqrt{2\pi 250}\cdot e^{-250}250^{250})^4} \approx \frac{\sqrt{2\pi 1000}}{4\pi^2 250^2}(\frac{1000}{250})^{1000} \approx 4^{992}.$$

Just a slightly smaller number than before....

(3) Combinations

First, for the moment we consider an extension of the permutation problem. All the ways of assigning n labels to k (labeled) slots, not allowing for redundancy and assuming $n > k$. It should be obvious now that this is $\underbrace{n(n-1)(n-2)\cdots(n-k+1)}_{k \quad objects}$. That is, we assign n label to the first slot, then we have (n-1) labels for the next slot, etc. We will use the notation P(n,k) to denote $n(n-1)\cdots(n-k+1)$.

With combinations, we would like to assign n labels to k slots but we don't care about the order. That is, ABC and BAC are seen as identical objects. Now we have enough tools to figure this out. The number of ways of assigning *n* labels to *k* (labeled) slots is *P(n,k)* = $n(n-1)\cdots(n-k+1)$. For any given assignment of k (labeled) slots, we consider all ways of ordering the slots equivalent. There are *k!* ways of ordering the slots. So we have *P(n,k)/k*! combination of assigning *n* labels to *k* slots regardless of order. This formula can be written more neatly in using the factorial notation,

$C(n,k) = \frac{n!}{(n-k)!k!}$. (You should verify that n!/(n-k)! = n(n-1)..(n-k+1).)

If you are confused, one way is to simply memorize the phrase "n choose k = $C(n,k) = \frac{n!}{(n-k)!k!}$".

2. Probability

Finally, onward to probability....

Recall from above, the discussion on events and so on. Here I will summarize some of the things that come out in elementary probability textbooks. We will use the notation P(E) to denote the probability of the event E, e.g., three dots on top of a dice. Following tradition, we will use the symbol Ω to denote the total sample space. Therefore, P(Ω) = 1, by definition. If we are dealing with a random variable we might use a notation like P(X=1) to mean "the probability that the variable X takes on the value 1 for a trial". If you remember the definition of random variable, you will see that a notation like P(X) does not make sense because a random variable is not an event. When it takes on a specific value like X=1 (or a range of values like 0<X<1), it is an event. This notation is not to be confused with probability distribution functions discussed below.

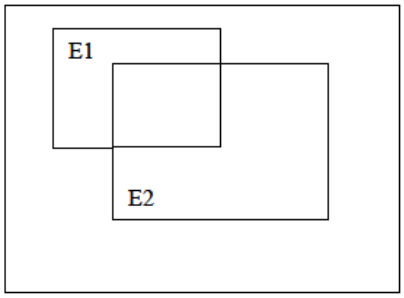
A few more notations are in order, but to do this we should examine the notion of events better (because this always confuses the hell out of me). Again using the dice example, the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. An elementary event from this space might be something like {3}, however, we can also consider $\{1,3\} = \{1\} \cup \{3\}$ an event of interest as well. That is, not only are the elements of the sample space events, but their unions and intersections are also events. For example, two kinds of events from Ω might be E1 = {1,3,5} (all odd numbers) and E2 = {1,2,3} (all numbers less than or equal to 3}. We can have these kinds of events:

$E3 = E1 \cup E2 = \{1, 2, 3, 5\}$
$E4 = E1 \cap E2 = \{1,3\}$
$E5 = E1 \cap \{4\} = \phi$
$E6 = \{1\} \cap \{2\} = \phi$
$E7 = \{1\} \cup \{2\} \cup \{3\} \cup \{4\} \cup \{5\} \cup \{6\} = \Omega$ (*)

The intersection of two (or more) events, like E1 and E2, is sometimes simply denoted E1E2. When the intersection of two events is the empty set, like E5 above, they are called mutually exclusive events. The union of all events is the sample space. Therefore, to be more rigorous, a sample space must be comprised of things such that we can take unions and intersections (actually, in a specific way but we won't care here). Now we have some rules for computing the probability of these events.

Summation rule

$P(E1 \cup E2) = P(E1) + P(E2) - P(E1E2)$ (Eq 1)



That is, the probability of the union of two events E1 and E2 is the sum of the probability of the each by themselves minus the probability of the intersection. This is because, in effect the probability of the intersection part is summed twice in the quantity P(E1)+P(E2) as shown in the left figure.

For mutually exclusive events, E1 and E2, we have $E1 \cap E2 = \phi$ and therefore P(E1E2) = 0, so

$P(E1 \cup E2) = P(E1) + P(E2)$ (Eq 2)

Suppose we have a series of mutually exclusive events $E_1, E_2,,,E_n,,$ .Then we have,

$$P(\bigcup_{i=0}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)\,.$$ (Eq 3)

Here we are simply extending the Eq 2 to many such events.

If in particular the union of all the mutually exclusive events is Ω, then we have,

$$P(\bigcup_{i=0}^{\infty} E_i) = P(\Omega) = \sum_{i=1}^{\infty} P(E_i) = 1. \qquad \text{(Eq 4)}$$

Now we have made more precise the statement in the beginning about the probability of the events in the sample space summing to one. The events must be mutually exclusive and the union of those events must be Ω.
Some algebraic rules for probabilities are in order.

P(E1 ∪ E2 ∪ E3) = P((E1∪E2)∪E3)
P((E1∪E2)E3) = P(E1E3∪E2E3)
P(E1E2E2) = P(E1E2)

These are, of course, just set operations.

For the most part, many questions can be answered without resorting to the calculus of probability. However, sometime we can get awfully confused as to things like when to sum, when to subtract, when to multiply and so on. It is useful to be able to get back to these basics at those times.

Conditional Probability

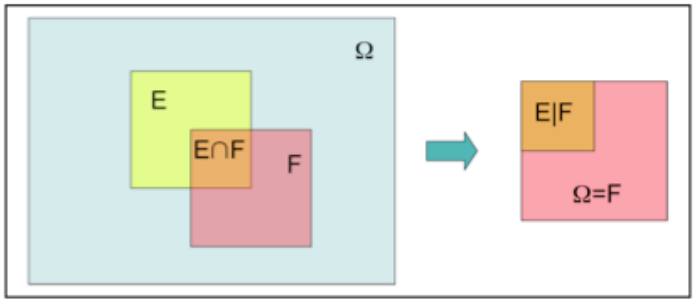
The meaning of conditional probability is fairly clear. We might toss a dice twice and be interested in a question like, what is the probability that the sum of the two results is 6, given that the first dice toss was 3? Such conditional events are denoted P(E\F) by which we mean the probability of the event E given F has been already realized. (The notation with the backslash "\" is also often substituted with just a bar "|". I will use the two interchangeably.) The formula is given as,

$$P(E \setminus F) = \frac{P(EF)}{P(F)} \qquad \text{(Eq 5)}$$

I always forget this formula, so it's good to remember why this holds. First, the top part is obvious. When we say the event E given F, both of the events have to occur so we have the P(EF) term. For example, suppose we are considering the experiment of throwing a dice twice. If E = {{1,5},{2,4},{3,3},{4,2},{5,1}} and F = {{3,1},{3,2},{3,3},{3,4},{3,5},{3,6}} then the event corresponding to "E given F" is the event EF = {3,3}. (Why is this?)

Why divide by P(F)? This is because we assume that the event F has already happened, which then redefines the relevant sample space. In the previous example, we go from a

sample space with 36 elements to {{3,1},{3,2},{3,3},{3,4},{3,5},{3,6}}; that is, the space of the F event. Then, within the set of possibilities defined by the occurrence of F, we are looking for the particular outcome that also belongs to E. Therefore, a renormalization is in order so that the probability of the new sample space equals one. The figure below gives and geometric intuition about conditional probabilities.



A rearrangement of Eq 5 yields:

P(EF) = P(E\F)P(F) (Eq 6)

That is, the probability of E∩F is the probability of the event F and the conditional probability that E will happen given F has happened. This is pretty intuitive and easy to memorize. So you can try memorizing Eq 6 instead. However, it should be noted that the conditional probability is not defined if P(F) = 0, which is not obvious from Eq 6.

Independent events

Two events are defined as independent if

P(EF) = P(E)P(F). (Eq 7)

You should remember this definition along with the notion of events and such because in the course of doing science, there will be many opportunities to be totally confused about independence. The common example is the probability that two dice tosses will yield the combination {1,1}. We all know this is 1/6*1/6 = 1/36. But it is useful to think about what the events are. The event set relating to the case that the first dice is "1" is:

E = {{1,1},{1,2},{1,3},{1,4},{1,5},{1,6}}

The event set that the second dice is "1" is:

F = {{1,1},{2,1},{3,1},{4,1},{5,1},{6,1}}

Thus, EF = {{1,1}}. By counting the events in the total sample space (36 items) and assuming the equiprobability model, we have P(E) =1/6 and P(F) = 1/6 and P(EF) = 1/36. Thus, *if P(EF) = 1/36*, then P(EF) = P(E)P(F) and we come to the conclusion that E and F are independent. If P(EF) were to be something other than 1/36 (which we might estimate from doing the empirical experiment), then we would come to the conclusion that the two events are not independent. **Independence is different from correlation. Correlation is a kind of a statistic (which we will discuss later). Independence on the other hand is a property of the probability model.**

An example will show the importance of considering the events carefully. First, we extend the notion of independence to many events:

Events $E_1, E_2, ..E_n$ are independent if $P(E_1E_2..E_r)=P(E_1)P(E_2)..P(E_r)$ for any subset of the *n* events.

That is, the product rule must hold for any pair, any triplet, and so on for the *n* events. Now the example:

Let a ball be drawn from an urn containing four balls, numbered 1,2,3,4. Let E = {1,2}, F = {1,3}, and G = {1,4}. Then P(EF) =P({1}) =1/4 and P(E)P(F) = 1/2*1/2=1/4 so the event E and F are independent. You can check likewise that E and G are independent and G and F are also independent. But P(EGF) = P({1}) = 1/4 and P(E)P(G)P(F) = 1/8 so the events E,G, and F are not triplet independent. This would be quite confusing unless we carefully considered what the events and intersections of events look like.

Bayes' formula

The Bayes' formula is one of the most important formulas you will see in this course. It is also the basis of Bayesian statistics, which we will learn later.

Let E and F be some event and we will also let E^c and F^c denote the complement of E or F. That is, E^c means all the events in the sample space except E. We know that $P(E^c) = 1-P(E)$ by construction. Now, the event E can be expressed as

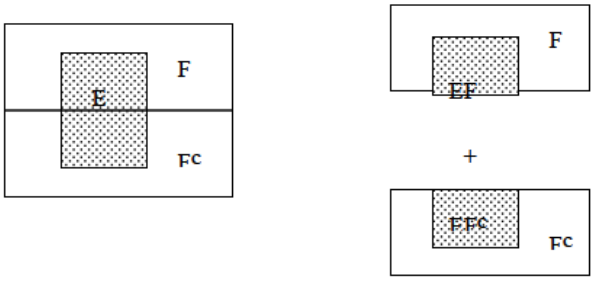
$E = EF \cup EF^c$ (why?)

so we have

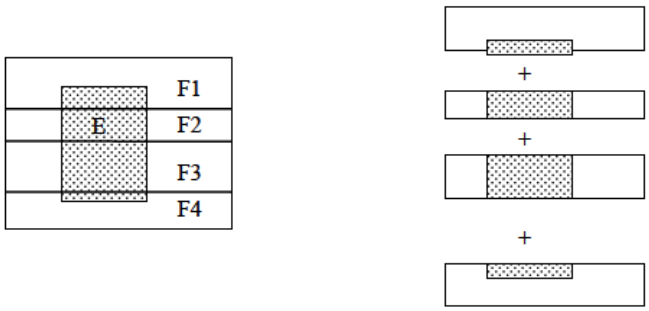
$$\begin{aligned} P(E) &= P(EF) + P(EF^c) \text{ (because these two events are mutually exclusive)} \\ &= P(E\backslash F)P(F) + P(E\backslash F^c)P(F^c) \text{ (from Eq 6)} \\ &= P(E\backslash F)P(F) + P(E\backslash F^c)(1-P(F)) \qquad \text{(Eq 8)} \end{aligned}$$

The interpretation of Eq 8 is that we are "partitioning" the probability of the event E with respect to another set of mutually exclusive events (F and F^c).

This can be seen on the right,



This kind of sectioning can be extended to any set of events as long as the set of events are disjoint (mutually exclusive). For example,



The corresponding equation to the picture would be,

$$P(E) = \sum_{i=1}^{n} P(E \setminus F_i)P(F_i) . \quad \text{(Eq 9)}$$

Sometimes this is called the law of total probability. It just says that if we have an event of interest like E, it can be calculated as the union of mutually exclusive conditional events.

Now we are ready to go on. We often have the question of the form "given event F what is the probability of the event E?". Sometimes these kinds of questions are more easily answered if we could somehow transpose from statements of the form "given E what is the probability of F?". The purpose of these manipulations is to try to do this. So writing the statement we have,

$$\begin{aligned} P(E \setminus F) &= \frac{P(EF)}{P(F)} \\ &= \frac{P(F \setminus E)P(E)}{P(F)} \\ &= \frac{P(F \setminus E)P(E)}{P(F \setminus E)P(E) + P(F \setminus E)P(E^c)} \end{aligned} \quad \text{(Eq 10)}$$

The first line is from Eq 5, the second line is from Eq 6 for the joint event EF. The last line is from the total probability formula, Eq 8.

More generally we have,

$$P(E_m \setminus F) = \frac{P(F \setminus E_m)P(E_m)}{\sum_{i=1}^{n} P(F \setminus E_i)P(E_i)} \quad \text{(Eq 11)}$$

This is known as Bayes' formula. The use of this is not obvious, however, note that we have changed all of the terms that look like P(E\F) on the left to P(F\E) on the right. One possible application of Bayes' formula is in questions like "Given some data X, what is the probability that a model Y is the true model?". That is, what is P(Y\X)? We usually know how to compute probability of the data given a model but not the other way around. Bayes' formula allows us to invert the P(X\Y) to P(Y\X). The top part of Bayes' formula is P(X\Y)P(Y) which can be interpreted as the conditional probability of the data X given the model Y times the unconditional probability of the model. The P(Y) terms are called prior probabilities and there are reasonable ways to obtain this number. Similar interpretations are possible for the bottom part.

3. Random variables

As discussed initially, random variables are real numbers that are labels for stochastic events. Most probability modelings are done with random variables rather than the actual event. An example of random variable is a variable X that takes the values 1,2,3,4,5,6 depending on a dice toss. We do the obvious assignment and say X = 1 when one dot turns up and so on. So far this isn't too useful. However, we can define random variables so that they will represent certain subsets of the sample space that are interesting. Consider the toss of two dices now. There are 36 possible elementary events. We might define a random variable X such that it is the sum of the number of dots on the top of the two dice. For example, X = 2 when we have the event {1,1}, X = 3 with the events {1,2} or {2,1}, and so on. A probability value can be attached to each value of the random. After a little elementary calculation we come up with P(X=2) = 1/36, P(X=3) = 2/36, P(X=4) = 3/36,…, P(X=6) = 5/36….etc. The set of values taken by the random variable and their probabilities is defined by something called the probability distribution of the random variable. (A more precise definition is needed but we will ignore it for the moment.)

Most of the "real" probability operations are done on random variables. After all, once we have the values of the random variable and their probabilities, we don't need to refer back to the original sample space. The sample space can be represented by (some combination of) the range of values of the random variable. However, it is important to remember that in principle there is supposed to be stochastic events and a probabilistic sample space "behind" the random variables. In practice, we know a lot about certain random variables but not a lot about the real sample space and event probabilities. For example, the probability distribution of protein 3D geometry is often modeled by a formula that looks like:

$$P(g) = Ce^{-\Delta(g)/T} \qquad \text{(Eq 12)}$$

where g is a particular geometry and $-\Delta(g)$ is something called the free energy of the geometry. Eq 12 is a form of an "exponential random variable", whose mathematical and probability properties that we know a lot about. But, we don't really know much about the underlying biophysics of individual protein fold configurations.

We often choose a random variable with a certain distribution to represent the stochastic phenomenon of interest; assuming that the random variable and its probability distribution is in fact what we would have ended up with had we computed the event probabilities from a detailed model. This is what we mean by probabilistic modeling.

Ok, now some random variables. I will only cover a very small set. You should look up things in a probability book. Familiarity with different random variable distributions is very important for modeling. As stated above, the random variables are characterized by their probability distribution. We will make this notion more precise. The random variables may take discrete values or continuous values. The discrete random variable is characterized by the **probability mass function** (*pmf)*. The values of the *pmf* tell us the probability that the random variable will take a particular value. For example, if we construct a random variable, X, for a single dice toss, we find its *pmf* as, p(1) = 1/6, p(2) = 1/6, p(3) = 1/6, p(4) = 1/6, p(5) = 1/6, p(6) = 1/6 under the equiprobable model.

The interpretation of a *pmf*, *p(z)*, is Prob(X = z) = *p(z)*. This seems like a lot of notation nonsense. But remember that X is the random variable that can take many different values by chance. The *pmf* is a function whose argument, *z*, does not change by chance.

If the random variable is continuous then we have a function analogous to pmf called the **probability density function** (*pdf*; often *pmf* is also called *pdf*). There is now an important distinction between *pmf* of a discrete random variable and the *pdf* of a continuous one. Suppose *f(z)* is a *pdf*. The value given by the *pdf* function, say *f(1),* does not represent the probability of X=1. In fact, the probability of a particular single number for a continuous random variable is not defined in an intuitive manner (for technical reasons we won't discuss here). Rather, the *pdf* is a function whose ***integration under the curve*** gives probabilities of a range of events. This is a bit technical and I will avoid going into the full detail here, again you should look up an elementary probability book for the full details.

For both the *pdf* and *pmf* we define yet another function called the **cumulative distribution function** (*cdf*) or sometime just called the probability distribution function. This is a function that gives us the probability that the random variable is less than or equal to some value. So if we denote *F(z)* as the *cdf* for some random variable,

$$F(z) = \mathrm{Prob}\{X \le z\} \quad \text{(Eq 13)}$$

If the random variable is a discrete variable and *p(x)* is the *pdf*, $F(z) = \sum_{x=-\infty}^{z} p(x)$. The summation proceeds from negative infinity or the smallest value that the random variable X is defined on up to the point *z*. Similarly, for continuous random variables we have $F(z) = \int_{-\infty}^{z} p(x)\,dx$. The *pdf, pmf*, and *cdf* all have the kind of properties that you think they should have. That is, *pdf, pmf* ≥ 0, and $F(\infty) = 1$.

Everything about the random variable is characterized by its *cdf* (or implicitly by the *pdf* or the *pmf*). We will examine a small number of well-used random variables.

(1) Bernoulli random variable

This is one of the simplest random variables. Suppose that the sample space for a trial is {success, failure}. We will represent the sample space by the random variable X and set X = 1 for the event {success} and X= 0 for the event {failure}. We assume that there is some parameter $0 \geq p \geq 1$ that denotes the probability of success. We now have the *pmf*,

$p(0) = 1-p$
$p(1) = p$

(2) Binomial random variable

Suppose we now have *n* independent Bernoulli trials, the binomial random variable represents the number of successes out of *n* trials. The *pmf* is,

$$p(x) = C(n,x)p^x(1-p)^{n-x}.$$

The function *C(n,x)* is the combinatorial function from above.

(3) Multinomial random variable.

We extend the notion of the binomial random variable to when the outcome of each trial might be more than two events. That is, instead of {success} and {failure} type of outcomes we might have $\{E_1, E_2,,,E_m\}$ possible outcomes. For example, a gene with *m* different alleles. We denote the probability of each outcome for a given trial as p_1, $p_2,,,p_m$, respectively. After *n* trials we would like to know the probability that we have say, f_1 of the E_1 event, f_2 of the E_2 event, and so on. We will use the notation $p(f_1, f_2,,f_m)$ to mean this probability. Then we have,

$$p(f_1,f_2,,f_m) = \frac{n!}{f_1!\,f_2!..f_m!} p_1^{f_1} p_2^{f_2} \cdots p_m^{f_{m1}}$$

The above formula assumes that $\sum_{i=1}^{m} p_i = 1$ and $\sum_{i=1}^{m} f_i = n$.

(4) Poisson random variable

A random variable taking the values 0, 1, 2...is called a Poisson random variable if the *pmf* is

$$p(x) = e^{-\lambda}\frac{\lambda^x}{x!}$$

(See how now we are simply defining the random variable just through its *pmf*?)

Now we look at some continuous random variables.

(5) Uniform random variable

Well can't leave this one out. It is,

$$p(x) = \begin{cases} 1 & if \quad 0 < x < 1 \\ 0 & elsewhere \end{cases}$$

(What does this mean in terms of probability models?)

(6) Exponential random variable

Defined by the *pdf*,

$$p(x) = \begin{cases} \lambda e^{-\lambda x} & if \quad x \geq 0 \\ 0 & if \quad x < 0 \end{cases}$$

(7) A Gaussian (= normal) random variable.

Defined by the pdf,

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

4. Fun things with random variables

Here are some properties.

- The sum of a two (or more) random variables is another random variable.

For example, let's let Z = X+Y where X and Y are two independent Bernoulli random variables, both with the parameters *p=1/2*. The value that Z can take is obviously {0,1,2}. The *pmf* of Z is, *p(0) = ¼* , *p(1) = ½* , and *p(2) = ¼*, which you can calculate with just a little effort.

Let Z be the sum of two independent discrete random variables X and Y. Denote the *pmf* of X and Y by *f(x)* and *g(x)* respectively. To obtain the *pdf* of Z = X+Y, we go through the following. For each possible value of Z, say Z = *v*, we need to compute the ways in which X+Y can sum up to *v* and then compute their probabilities. First start from the

lowest value of X that is smaller or equal to *v*, say *w*. We know this probability as *f(w)*. Now we find the probability of Y = *v-w* which is *g(v-w)*. Since X and Y are independent we have the probability *f(w)g(v-w)* for this combination. Next, we change the value of X, say to *w'*, then find *g(v-w')* add *f(w')g(v-w')* to *f(w)g(v-w)* (because these two are mutually exclusive events). We continue this until we have exhausted the possible combination in which X+Y = *v*. (This computation is called finding the convolution of two distributions.) Sounds tedious and it often is. We already did one example above. The binomial random variable is the sum of *n* independent Bernoulli random variables. If X and Y are continuous random variables, we do analogous computations using integrals instead of sums.

- The sum of a random variable and a constant is also a random variable.

Some of you may remember the good old ANOVA model Y = X + δ. Here we are saying that the dependent variable Y is a random variable that is the sum of the random variable δ and a constant X (but with different values for different effects). Of course, the sum of a random variable and a regular variable is also a random variable (like the regression model). A function of a random variable, say *f(X)* is also a random variable. For example, $f(X) = X^2$. (Technically, one has to be careful with the idea of a function here since random variables are not the usual kinds of numbers. We will ignore things like that here, but may do an example in class where this distinction is important.)

We define a quantity that we call the expectation of a random variable by,

$$E[X] = \sum_{x:p(x)>0} xp(x) \quad or \quad \int_{x:p(x)>0} xp(x)\,dx$$

The first definition is for discrete *pmf* and the second is for continuous *pdf*. The notation *x: p(x)>0* means to sum or integrate only over those values of *p(x)* that are greater than zero.

We can also take the expectation of a function of a random variable, say *E[f(X)]*. We have,

$$E[f(X)] = \sum f(x)p(x) \quad or \quad \int f(x)p(x)\,dx$$

That the expectation of a function of a random variable looks like this requires a proof, which we won't be concerned with. An example of the expectation of a function of a random variable is what we call the variance of the random variable. You are familiar with the formula

$$\mathrm{var}(X) = \sum (x - E(x))^2 p(x) .$$

As you can see, we can regard the function to be $f(X) = (x-E(x))^2$.

Finally onto

Statistics

Statistics is where probability meets data. The basic idea is simple. First, represent the data by some number—we call this number a "statistic". Model a probability distribution for this number by associating it with a random variable. Use the probability distribution to make probabilistic statements about the data and test model of the data. We look at each step in turn.

"Statistic"

As mentioned, a statistic is a number or a set of numbers computed from measured data that is supposed to capture an empirical phenomenon in a meaningful way. For example, "body weight" is a statistic that captures some notion of obesity. Of course, once you start thinking about the problem you realize that this is not such a good measure since a person might be thin and tall and therefore heavy. So, the statistic might be revised to weight/height. This does a better job of capturing the idea of obesity. Some more careful thought results in weight/height^2, which is called the Body Mass Index (BMI). (Why should we take the square of height?) The key is that the statistic is a representation of the empirical situation and it should mean something to the person who sees the statistic—like BMI large is bad, BMI small is bad, somewhere in the middle is good.

There is no algorithm for deciding on a good statistic. Finding a good statistic is in fact part of the art of modeling. A more general notion of statistic is called "feature space" in the machine learning literature. A feature space is a collection of numbers (thus a "space") that represents the relevant "features" of the empirical data. While there is no good algorithm for deciding on a statistic or a feature, there are some things to consider.

Guide points to constructing your own statistic.

1. Imagine the extremes of the situation that you want to characterize. In the obesity example, this is some notion of normal bodies and abnormal bodies. Construct a number that we can compute from the data such that it will take high (or low) value for one extreme and low (or high) value for the other extreme. (What should we do if there are more than two kinds of extremes?)

2. Imagine intermediate situations and try to verify that as we transit from one extreme situation to the other, the value of the statistic will not jump around wildly. Our intuition demands that the values of the statistic mean similar things at all the value intervals.

3. Try to normalize the statistic. It is difficult to understand a number if it is unbounded (e.g., can reach infinite values). Also, try to see if there are other kinds of modulating

factors to consider. For example, one might try to include in the BMI an adjustment for age or ethnicity.

4. See if we can compute the probability distribution of the statistic under a reasonable stochastic model. (See below.)

4. See if we can construct scenarios in which the statistic will misbehave such as different sample sizes, singular situations (e.g., what if somebody had a large tumor?), etc.

Computing probabilities of the statistic

Once we have a statistic that represents the empirical data, we would like to convert that into probability statements. The reason to do this is two fold. First, we have a better intuitive feel for probabilities than for the original numbers. I don't really know what BMI = 18.5 means but I have a better idea if you tell me that the probability of BMI less than 18.5 for a random individual drawn from the US population is 0.15. Second, a probability statement is universal. A probability statement can be made about, say genomic DNA content and BMI, and they can be interpreted together. So, we might have the statement that the probability of an individual having particular nucleotide "A" at position 12762847 in chromosome 22 is 0.1, and the probability of the individual having BMI < 18.5 = 0.15 and be able to talk about the joint probability of two events using standard probability calculus described above.

Unfortunately, generating a probability for a statistic is often very difficult. First, you have to come up with a stochastic model that can be applied to the statistic. Then, you have to be able to compute the probability using that model.

The most straightforward stochastic model for a statistic is to claim that the statistic is a random variable from some known probability distribution. This probability distribution might be mathematically described, such as the Gaussian distribution, or described by an idealized "population" (e.g., the population of US) for which the random variable is obtained as a random sample from the population. Most of the statistical applications use this method of modeling. Thus, we might say that the BMI is a random distribution from a normal distribution with the parameters: mean = X and variance = Y. How do we know that this is a reasonable model? There isn't always a good way to know. Like any inductive inference, the basic way is to get lots of examples and then show that the distribution of those examples "looks like" the normal distribution. There are more rigorous ways to measure fit of the data to the idealized model, some of which we will discuss in class. How do we know which model to propose? This requires familiarity with the various standard models, i.e., the well-characterized probability distributions, and a lot of experience.

The other way of constructing a stochastic model for your statistic is to build it up from some more elementary probability models. For example, we might have a model for nucleotide identity at some position that says, we start from some ancestral identity, say "A", and this nucleotide has the possibility of mutating to "C" each generation with the

probability 0.00000004, and the current statistic is measured on a sample drawn after 10,000 generations. This kind of model creates two problems. First, we still have to find ways to make sure the more elementary probability model is reasonable, like the probability of mutation. Second, we have to know how to calculate the probability we want from the elementary models. The second part can be very involved and often requires deep mathematical statistics expertise.

Estimating model parameters

In modeling, when we construct and choose a model we almost never deal with a singular model, but a family of models. In fact, this is probably true for any theory or explanation that we normally use. Suppose we came into a room and found a dead body. We might have the theory "there was a murder". A "murder" is a generic explanation of what might have happened. To apply this theory to our particular case we then need to add specifics such as the identity of the victim, the murder weapon, the time of death, the mode of death, etc. "Murder" is our model or theory and things like "murder weapon" are the **parameters of the model** and things like "the blunt end of a heavy statistics textbook" are the value of the parameter. The model sets the general structure of the explanation and the parameters are factors that may vary within the model (to fit a particular situation).

Going back to the model of BMI, if we assume that a normal distribution is a good model for the BMI statistic then we need two parameters, the mean and the variance. Sometimes we have a priori knowledge about the parameters and we are done with that. But, most of the time we have to estimate the parameters from the data themselves. The general principle that we use is to "find the model parameters that best fits the observed data". That is, we try to give the model the best chance as possible to explain the data—if that doesn't do a good job then we reject the model (see below).

Let's re-examine what we have discussed. If somebody gives us a bunch of information, we initially have no idea what we can say about it at all. The information or the data could have come from anywhere. Given some organization and perhaps other hints, we decide on a family of models as a reasonable explanation. But, since this is a family of models, we want to narrow it down to a specific model--so we examine the data and estimate the model parameters from the data. We then end up with our best specific explanation of the given information/data. The process of choosing a suitable family of models and specifying the parameter values is called ***Statistical Inference***.

Testing Hypotheses

Using the statistical techniques discussed so far we can organize and summarize information and hopefully attach a model-based explanation. We have not yet discussed the most fun part of statistics--testing hypotheses. But, before discussing formal hypothesis testing lets prime ourselves by considering how statistics allows us to ask and answer useful questions. Consider how many times we ask the question "Is this unusual?". For example, "I saw a Rolls Royce this morning; is that unusual?", or "I did

some crosses with a mutant strain of peaplants and found a ratio of 123:110 for the green and yellow albumen; is this strange?", or "I found a gene in a T-4 phage with a GC content of 68%; could this be a host gene?". Standard hypothesis testing involves setting up some expectations than looking for the unusual events that might invalidate my hypothesis.

As an example, we might have the belief that genes are constantly mutating resulting in sequence differences between divergent organisms, say humans and mouse. We now have two genomes and our ***hypothesis*** is: "No stretch of the mouse and human genomes are exactly identical for more than 100 base pairs (bp)". We ***test*** this hypothesis by sequencing the genomes and looking at 100 bp pieces. Every time we see two different 100 bp pieces from human and mouse we gain inductive support for our hypothesis. (Do you know what is meant by inductive support? See K. Popper.) If we see two 100 bp pieces that are exactly identical, we reject our hypothesis. That is, the event "two pieces of 100 bp genome substring in chromosome 22 that is identical between humans and mouse" belongs a set of events (other events in this set might other identical pieces in other chromosomes) that forms a ***rejection*** of the hypothesis.

Now, in this setup we had the positive belief "No stretch of the mouse and human genomes are exactly identical for more than 100 base pairs (bp)". This means a sighting of identical 100 bp sequences is a definite rejection of the hypothesis. The world at large is rarely this clear cut. It is most likely that a more reasonable scenario is "we are *unlikely* to see 100 bp identical pieces".

Statistical hypothesis testing gives us a way to formalize the problem of probabilistic hypothesis tests. First, the hypothesis being tested (called the null hypothesis) is set up. Then, we attach a reasonable probability model to the hypothesis (a more strict interpretation would be that the probability model itself is the hypothesis being tested). For example, a precise probability model version of the previous scenario might be to detail a model that give me the probability of seeing identical stretches of human-mouse genomes longer than length k.

Typically we are not really interested in the probability hypothesis itself (i.e., that the probability of identical genomic pieces of length $> k$ is given by some function). We are actually interested in the probability model as it is computed from some biological process model. An example of such a biological process model might be what is called the "neutral drift model of sequence evolution." This neutral drift model of sequence evolution might detail various things so as to give us a probability function that yields the probability of identical sequences of length k.

Now, suppose we have two sequenced genomes and we find a stretch of identical sequence that is 1000 bp long. My null hypothesis (or the null model) tells me that the probability of such an identical stretch of the genome is 0.000001. Then I might reject the hypothesis of "Neutral drift model of sequence evolution". Of course, when we make this argument we are agreeing to accept the danger that, in fact, the two genomes have evolved by a neutral drift process and what we are seeing is just a rare chance event. This

"rare event" happens with probability 0.000001, so we are accepting the 1 in a million risk that the null model is, in fact, correct. We then say "we reject the hypothesis of the neutral drift evolution at the 0.000001 significance level".

There are several important points to keep in mind in this setup. First, we may choose to reject the null hypothesis at any kind of significance level—this is a matter of how risk adverse we want to be. The standard is to choose something like 0.05. This means that we are allowing ourselves to falsely reject the null hypothesis 5% of the time, when in fact the null hypothesis is true. We might modulate this significance level dependent on what we want to do next. Second, the rejection events that sum up to a certain significance level can be chosen by the investigator in many different ways. We will discuss this point in class. Finally, when we reject a null hypothesis, we are rejecting the probability model of the null hypothesis. Strictly speaking, we are not rejecting the biological model underlying the probability model. So, we might reject the neutral-drift model of sequence evolution not because that is not how the biology happens but because our probability model of the neutral-drift model might be wrong. For example, we might have the wrong parameters for the model and it may turn out that the mutation rate in mouse and humans were much lower than what we originally thought. We need to keep both of these possibilities in mind when dealing with empirical data.