Lecture 2: Probability and Some Random Variables

"Randomness scares people."

-- Fran Leibowitz

1. A Review of Set Theory and Probability

Probability is a generalization of the idea of weight or area or volume. There's a universal set of objects denoted Ω ('omega') and it has a total mass, area, volume of 1. Each object has a particular mass, etc. and we're interested in figuring out the mass, etc, of any set *A* of objects we select out of Ω. It's evident that the weight of *A* is just the weight of the objects it contains denoted P(*A*). The weight of the set formed out of all the sets not in *A* (denoted $\bar{A}$, "A complement") is obviously 1-P(*A*):

$$P(\bar{A}) = 1 - P(A) \tag{1.1}$$

The empty set $\varnothing$ contains no objects and has mass, etc. of 0. It's the complement of Ω and we could compute its mass from Eq. 1.1 if it weren't obvious. If sets A and B have no objects in common, then their *union* $A \cup B$ has a mass that is the sum of the masses of *A* and of *B*. We can write,

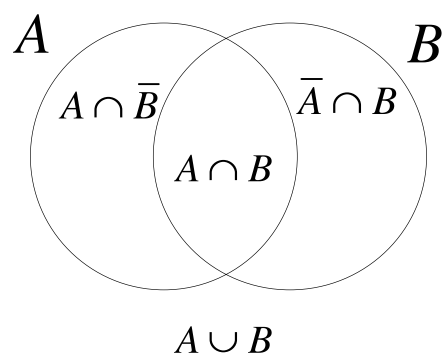
$$P(A \cup B) = P(A) + P(B), \qquad A \cap B = \varnothing \tag{1.2}$$

where $A \cap B$ denotes the *intersection* of *A* and *B*, the objects they have in common. We can also show that,

$$P(A \cup B) = P(A) + P(B) - P\left(A \cap B\right) \tag{1.3}$$

where the last term compensates for double-counting objects that are in both *A* and *B*.

If there are four objects total, then there are 16 possible subsets. Assigning non-negative numbers that sum to less than 1 to any three of these specifies a probability measure on all 16 subsets and that's usually how a discrete probability measure is usually specified by tabulating the probabilities of its 'objects' (usually referred to as 'atoms'). If there are *n* atoms, then there are 2^n possible subsets (why?) and assigning *n-1* non-negative



numbers to *n-1* of the atoms specifies a probability measure on all the subsets so long as the *n-1* non-negative numbers sum to less than or equal to 1.

2. **The Borel Sets and the Cumulative Distribution Function**

In the last lecture, I claimed that the cdf (cumulative distribution function) of a real-valued random variable completely specified the random variable. The cdf of a random variable is, $F_X(x) = P[X \le x]$ and it may not be completely obvious to you that this expression tells us everything we might want to know about *X*. For example, suppose *X* is a discrete random variable. We might want to know the probability that *X* is 2. Or, if *X* is a continuous random variable, we might want to know the probability that it falls in the interval (0,1). At first glance, all the definition above tells us is the probability that *X* falls into intervals of the form $(-\infty, x]$:

$$F_X(x) = P\left[X \in (-\infty, x]\right] \qquad (1.4)$$

To compute the probability that the random variable *X* falls into other subsets of the real numbers, **R**, we need the laws of probability. I'll mention these in class, and you can look them up in Wakerley. In this section, I'll 'exercise' them a bit to figure out exactly what the cdf of a function tells us about its random variable.

The probability that a real-valued random variable is in a subset of the real numbers is only defined for some of the subsets of the real numbers that are called the *Borel sets* of the reals. They include all of the sets you'll ever be interested in and they're constructed recursively. The Borel sets include,

a. the empty set, $\varnothing$.
b. the set of real numbers, R.
c. all the half-open intervals of the form $(-\infty, x]$,
d. and much more!

The remaining Borel sets are generated recursively by the following rules.

R1. If *B* is a Borel set, then so is its complement $\bar{B}$, the set of real numbers not in B.

R2. If $B_1, B_2, \cdots, B_n, \cdots$ are all Borel sets, then so is $\bigcup_{i=1}^{\infty} B_i$ (set union).

R3. If $B_1, B_2, \cdots, B_n, \cdots$ are all Borel sets, then so is $\bigcap_{i=1}^{\infty} B_i$ (set intersection).

It helps to think of the Borel sets as the result of a process somewhat akin to grinding sausage meat. Throw in all the open and closed intervals. Then turn the crank by following the following three steps.

1. Take all the complements of all the sets and include them as well.

2. Take all the (countable) unions of the sets possible. Include them too.
3. Take all the (countable) intersections of the sets possible. Include them too.

Repeat 1-3 over and over forever.

We can easily show that familiar subsets of the reals are Borel. By R1, all the sets of the form (a,∞) are Borel. And, consequently, so are all of the sets of the form,

$$(a,b] \quad = (-\infty,b] \cap (a,\infty),$$

using R3. And so are all of the singleton sets,

$$\{b\} \quad = \bigcap_{n=1}^{\infty}\left(b-\frac{1}{n},b\right]$$

by another application of R3. So are all of the half-open intervals of the form [a,b), the open intervals (a,b), and the closed intervals [a,b]. You will not be able to come up with an example of a subset of the reals that is *not* a Borel set so any question you ask about the probability that a random variable is in a specified subset of the reals will concern itself with a Borel subset. It's perhaps a little surprising but of no real interest to us that there are sets of the reals that aren't Borel sets. The proof is difficult and involves showing that there are 'more' subsets of the reals than there are Borel sets.

Now, as we construct the Borel subsets, we can also work out the probability that a particular cdf $F_X(x)$ assigns to each one. I'll go over some of this in class, but you might want to review what you know about probability (check with Wackerley) by verifying that it is so.

2. Cumulative Distribution Functions

So, the cumulative distribution function is just a compact way to specify the probability that a random variable is any one of the Borel subsets of the reals. If we idn't know about cdf's then we might have gone about pinning down a random variable by specifying it on some nice group of sets (like the intervals (*a,b*) and extending the definition to the other Borel sets). This way of defining probabilities is the standard way and we'll come back to it when we talk about random variables whose values are not numbers but rather realizations of stochastic processes. There won't be any equivalent of a cdf for these sorts of random variables but the basic idea of the Borel set construction will generalize very nicely.

Next we concentrate on the properties that a function must have to be a cumulative distribution function. The following theorem characterizes all the possible cumulative distribution functions (and, of course, all the possible real-valued random variables).

Theorem: $F: R \to [0,1]$ is a *cumulative distribution function* if and only if it satisfies

(i) F is non-decreasing.
(ii) F is right-continuous.
(iii) $\lim_{x \to \infty} F(x) = 1$.
(iv) $\lim_{x \to -\infty} F(x) = 0$.

The condition that *F* be right-continuous concerns step discontinuities in F. Intuitively, if *F* jumps at *x*, then the value of *F(x)* is consistent with the values of *F(.)* to the right of x. There is a certain arbitrariness to the requirement of right-continuity for cumulative distribution functions; if we substituted `left-continuity' (definition obvious) for `right-continuity' we could develop a perfectly good theory of random variables and everything would work out fine. Some (older) statistics books in fact use random variables whose cumulative distribution functions are left-continuous. It's only important that we be consistent.

Now, we need to develop some intuitions concerning cumulative distribution functions. The first concerns the `non-decreasing' clause in the definition. `Non-decreasing' does not mean `increasing.' A `non-decreasing' function may coast along at a fixed *y*-value for a while. Step functions whose successive steps have increasing *y* coordinates are examples of functions that are `non-decreasing' but not `increasing' over the duration of the flat tops of the steps. And here is the intuition: When a cumulative distribution function is constant (`non-decreasing') over an interval (a,b], the probability of that interval is zero. For example, give the step function shown in Figure 1, the corresponding random variable (call it B) can only take on the values 0 or 1. The probability of B being in any interval not containing 0 or 1 is 0.

The cumulative distribution functions that are (right-continuous!) step functions correspond to random variables that are discrete. To specify such a function we need only list the values at which the steps occur and the height of each step. The height of each step is exactly the probability with which the discrete random variable will take on the value at which the step occurred. This list, (value, height), is how a discrete variable is typically presented in an introductory course.

3. Some Discrete Distributions

The random variable X has a *Bernoulli distribution* with parameter *p* iff *P [X=0] = p* and *P [X=1] = 1-p*. We introduced them in the previous lecture in one of the examples. Bernoulli variables are excellent models of discrete choices and can be used to define more complex random variables defined in `stages.' In class, I'll describe a simple model of the detection process ('high threshold theory') in terms of Bernoulli variables.

A second discrete random variable is the *Geometric*:

$$P[X=n] = (1-p)^{n-1} p, \qquad n = 1,2,3,\cdots \tag{1.5}$$

which is a good model for processes with repeated identical trials that terminates as soon as a trial is 'successful.' Russian Roulette is such a process if we spin the cylinder of the revolver vigorously between trials. The Geometric distribution has a single parameter *p*.

The Poisson counting variable distribution is

$$P[X=n] = e^{-\lambda}\frac{\lambda^n}{n!}, \qquad n = 1,2,3,\cdots \tag{1.6}$$

There also a famous stochastic process called the *Poisson process* that we'll encounter later in the course. The term 'Poisson distribution' can be ambiguous, and we'll emphasize the 'counting variable' part of the name to aovid ambiguity. The Poisson distribution is an important model of many physical processes such as radioactive decay and it has several important mathematical properties we will return to in a later lecture.

In class, we will verify that the probabilities of the last two discrete distributions sum to 1. Next, let's develop Dirac notation to be able to write the pdf of discrete random variables such as the Poisson. Let $\delta_a(x)$ denote a Dirac function whose 'spike' is at location *a*. Then, let

$$p_n = P[X=n] = e^{-\lambda}\frac{\lambda^n}{n!}, \qquad n = 1,2,3,\cdots \tag{1.7}$$

for convenience. We want to put Dirac functions of height p_n at location *n*, giving the pdf,

$$\sum_{i=1}^{\infty} p_n \delta_n(x) \tag{1.8}$$

For the Bernoulli, the pdf is just $(1-p)\delta_0(x) + p\delta_1(x)$.

4. Continuous Random Variables

A second important class of cumulative distribution functions are those that are continuous and have a continuous derivative except at a finite number of points (`piecewise smooth'). These correspond to the continuous random variables. As the definition hints, the derivative of the cumulative distribution function (the probability density function pdf) plays an important role in characterizing them.

For any continuous random variable, the probability of getting any specific number is 0 and we can show that directly from the definition of the cdf. That means that the probability that a continuous random variable is in an interval doesn't depend on whether it is half-open, open, or closed.

We'll see in the next few lectures, that we often find ourselves combining discrete and random variables so that the resulting rv's are neither continuous nor discrete, but rather *mixed*. Intuitively, a mixed random variable has a cdf that occasionally jumps but is not constant between jumps.

Last time, we considered two continuous random variables, the Uniform U(0,1) and the Exponential with rate parameter λ. The exponential is really a family with a different distribution for each value of $\lambda > 0$.

It is natural to generalize the uniform distribution to the *uniform family* U(a,b) with pdf defined by,

$$f_{U(a,b)}(x) \quad = \begin{cases} \dfrac{1}{b-a} & a<x<b, \\ 0 & otherwise \end{cases} \tag{1.9}$$

A third example is the Gaussian family with parameters (μ, σ^2) whose pdf is,

$$f_{\mu,\sigma^2}(x) \quad = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1.10}$$

6. Expectation

In building models, we will often combine and transform random variables. The expectation of a function *h(X)* of a continuous random variable with pdf *f(x)* is

$$E(h(X)) \quad = \int h(x) f(x) dx \tag{1.11}$$

Suppose that we have, instead, a discrete random variable with pdf, in Dirac notation,

$$f(x) \quad = \sum_{i=1}^{\infty} p_i \delta_{x_i}(x) \tag{1.12}$$

and we would like Eq. 1.8 to 'work' even when we substitute a pdf of the form shown in Eq. 1.9 into it. The key turns out to be to define,

$$\int \delta_a(x) dx \quad = 1 \tag{1.13}$$

for any *a*. That is, a Dirac has unit area hidden inside its 'spike' and, when we integrate over a Dirac, we just add one. The usual properties of the integral give us the rest:

$$E\left(h\left(X\right)\right) \;=\int h\left(x\right) f\left(x\right) dx \;=\sum_{n=1}^{\infty} h\left(x_n\right) p_n \int \delta_{x_n}\left(x\right) dx \tag{1.14}$$

which simplifies to

$$E\left(h\left(X\right)\right) \;=\int h\left(x\right) f\left(x\right) dx \;=\sum_{n=1}^{\infty} h\left(x_n\right) p_n \tag{1.15}$$

and you likely recognize the right-hand side as the usual definition of the expected value of a discrete random variable.
We can extend Eq. 1.8 to mixed random variables by just breaking them into an ordinary pdf and the sum of a bunch of Dirac functions. (There's a little more to the story than this, but, for our purposes, this is all that's needed).

When $h(X) = X$, the resulting expected value is called the *mean.* When $h(X) = X^n$, the resulting expected value is called the n^{th} *raw moment* or n^{th} *moment about 0*. It is often denoted μ_n. When $h(X) = \left(X - \mu_1\right)^n$, the resulting expected value is called the n^{th} *central moment* where μ_1 is just the mean. The second central moment is the variance and, in previous courses, you may have hear people talk about higher central moments such as *skewness* (the 3$^{\text{rd}}$) and *kurtosis* (4$^{\text{th}}$).

This sort of analysis by moments does provide useful descriptive information about random variables but is no longer very important within statistics. We will need the terminology of moments to understand the *moment generating function* in the next lecture.