Classroom Notes

Math 394B Summer 2005

Week 5

1 Discrete Random Variables

In an operational sense, all random variables (R.V.) are discrete (and finite). This is in the same sense as, operationally, we can only deal with rational numbers (in fact, "real" numbers are defined only through limit operations on rational numbers), but it would be extremely cumbersome to perform any advanced calculations if we restricted ourselves that way.

A R.V. can be viewed as the result of a specific measurement which depends on the result of a random experiment - a function

$$X : \Omega \mapsto \mathbb{R}$$

Since our measuring instruments all have a finite range (no "infinite" readings - in fact, our readings are bounded by the scale of the instruments), and a finite precision (e.g., with a precision chronometer, in track and field events, the lower threshold for precision is currently $\frac{1}{100}s$), any measurement will produce numbers restricted within a finite set - possibly a huge set, but still finite (in the track and field example, we could only observe numbers $\frac{k}{100}$, where $k = 0, 1, 2, \ldots, N \times 3600 \times 100$, where N is the longest time (in hours) that the chronometer is able to report).

As noted, it would be very awkward to work within this frame, and we will indeed introduce infinite (better said, unbounded) R.V.'s, and *continuous* R.V.'s. It is good, though, to remember that what makes sense operationally are only discrete and finite approximations to these entities: all definitions should be constructed as limits from the discrete (and finite) setting in order to make sense unambiguously.

2 Describing Random Variables

We consider for the moment only discrete and finite R.V.'s. Let X be such a R.V. For any x, the function $p_X(x) = P[X = x]$ is well defined: it is equal to zero, except at a finite number of values, say $x_1, x_2, \ldots, x_n$, and

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

The function $p_X(x)$ is called the *distribution* or the *probability mass function* of X, and it carries all probabilistic information that depends on X (and on X only). Our job will be, in general, to compute probabilities of events that depend on what values X takes, and the distribution will be the basic tool.

2.1 Alternative Ways to Present Distributions

It is sometimes convenient to use alternative functions which carry the same information as $p_X(x)$. The prime example is

2.1.1 The Cumulative Distribution Function

This is also sometimes called the "Distribution Function", and is defined as

$$F_X(x) = P[X \leq x]$$

Clearly, if X takes on the values $x_1 < x_2 < \ldots < x_n$, and $x_k \leq x < x_k$,

$$F_X(x) = \sum_{x_j \leq x_k} p_X(x_j)$$

A few properties are easy to see:

1. $0 \leq F_X(x) \leq 1$
2. $F_X(x)$ is nondecreasing
3. $F_X(x)$ is constant except at points x_j, such that $p_X(x_j) > 0$, where it has a jump discontinuity equal to $p_X(x_j)$
4. Since we (arbitrarily) used a $\leq$ in the definition of F_X, instead of a $<$ sign, F_X is *continuous from the right*, i.e., for every x (whether it is a jump discontinuity point for F_X or not)

$$F_X(x^+) := \lim_{x \downarrow y} F_X(x) = F_X(y)$$

5. $\lim_{x \to -\infty} F_X(x) = 0$
6. $\lim_{x \to \infty} F_X(x) = 1$

By simple subtraction,

$$p_X(x_k) = F(x_k) - F_X(x_k^-)$$

(the last term is the left limit of F_X at x_k), so that knowledge of F_X allows us to calculate p_X, as well as vice-versa.

One "privilege" of the cumulative distribution function is that it is easily extended to the case of non-discrete R.V., and even allows to work on some difficult cases bypassing, at least initially, the necessary technical issues involved in advanced integration theory (more about this later)

2.1.2 The Survival Function

Obviously equivalent to F_X, this is a function especially popular in "survival analysis" (modeling the survival of organisms in biology and medicine), and "reliability theory" (modeling the same thing for machines):

$$R_X(x) := P[X > x] = 1 - F_X(x)$$

In the case of applications to survival analysis or reliability, X would be the "lifetime" of the object of study, and it would then be necessarily $X \geq 0$, but there is, of course, no reason to ban the use of R_X in more general situations.

You are urged to write up the properties of R that follow directly from the properties of F.

2.2 Parameters Of A Distribution

Before looking at yet more ways to characterize the distribution of a R.V., it is necessary to note that, often enough, it is pretty hard to come up with a full description of a R.V., and we may limit ourselves to limited information. That's why we introduce expected values, variances, etc.

2.2.1 Moments Of A Distribution

As discussed in the book, and in class, we are led to introduce the following numbers, associated to the distribution of a random variable:

- Absolute Moments (sometimes called "cumulants"):

$$m_k := E\left[X^k\right] \qquad k = 1, 2, \ldots$$

- Centered Moments (or just "moments")

$$M_1 = m_1$$

$$M_k = E\left[(X - M_1)^k\right] \qquad k = 2, 3, \ldots$$

The centered 2nd moment is known as the "Variance".

It is a boring, but straightforward calculation (using "Newton's Binomial Formula") to check that we can reconstruct absolute moments form their centered cousins and vice-versa.

The 1st and 2nd moments are the most commonly used, and it should be clear that they are very far from a full description of a R.V. Unless very specific assumptions are made, they do in fact carry extremely minimal information.

For a discrete and finite R.V. (as we are presently considering), if enough moments are given, we can reconstruct the distribution - but this problem (known

as the "moment problem") is not nearly as easy, or even solvable, in the general case. In fact, we note that

$$m_1 = E[X] = \sum_{i=1}^{n} x_i p_X(x_i)$$

$$m_k = \sum_{i=1}^{n} x_i^k p_X(x_i) \tag{1}$$

where $x_i, \quad i = 1, 2, \ldots n$ are the possible values of X.

Suppose we know the first $n-1$ moments. Then, together with the relation $\sum p_X(x_i) = 1$, (1) produces a system of n equations in the n unknowns $p_X(x_i)$. The coefficcient matrix of this system is

$$\begin{pmatrix} 1 & 1 & \ldots & 1 & 1 \\ x_1 & x_2 & \ldots & x_{n-1} & x_n \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ x_1^{n-2} & x_2^{n-2} & \ldots & x_{n-1}^{n-2} & x_n^{n-2} \\ x_1^{n-1} & x_2^{n-1} & \ldots & x_{n-1}^{n-1} & x_n^{n-1} \end{pmatrix}$$

The system does not have a unique solution if and only if there is a linear combination of the rows that is equal to zero, i.e. if

$$\sum_{k=0}^{n-1} a_k x_j^k = 0 \quad j = 1, 2, \ldots, n$$

i.e., if the n values taken by X happen to be n zeros of a polynomial of degree $n-1$ - and, since they are all distinct, they cannot be (there are, at most, $n-1$ such values). Of course, *solving* such a system, for n large, is a totally different and very long calculation.

2.2.2 Expectations Of Functions of a R.V.

The moments of X are a special case of expectation of a function of X: they are the expectations of the powers of X. More generally, we have seen that

$$E[f(X)] = \sum_{i=1}^{n} p_X(x_i) f(x_i)$$

(with the same notations as in (1)). The discussion of the "moment problem" suggests the further question: is there a class of functions f, such that knowledge of $E[f(X)]$ for all of them is enough to recover the distribution of X?

2.3 Generating Functions And All That

It should be clear that, in our case of finite discrete R.V.'s, there will be a lot of these classes. However, we might want to concentrate on classes that are easy to handle, and work to the same effect in more general cases.

2.3.1 The Moment Generating Function

It turns out that a useful class of functions is the family e^{tx}, of exponentials, with parameter t (each value of t gives a member of the family). The function

$$E\left[e^{tX}\right] = M_X(t)$$

is called the moment generating function, and the reason is that, by expanding it in powers of p, we find the absolute moments of X in the coefficients of the successive terms:

$$E\left[e^{tX}\right] = \sum_{k=1}^{\infty} E\left[X^k\right] \frac{t^k}{k!}$$

(in our finite case, there is no question of convergence, since $|X| \leq A$ for some number A). Clearly, just by referring to sec. 2.2.1, M_X will allow the reconstruction of p_X.

In the more general case of a continuous R.V., the moment generating function is known to analysts as the *Laplace transform* of the distribution of X. Of course, recovering the distribution form the MGF in a general setting means inverting a Laplace transform, which is not a task for the faint-hearted.

2.3.2 The Characteristic Function

This is the name that (unfortunately) is used in probability for what analysts call the *Fourier transform* of the distribution. That's why the function $1_A(x) = 1$ for $x \in A$, $1_A(x) = 0$ for $x \notin A$ is called the *indicator function* by probabilists, while in analysis it is usually called "characteristic function" (argh!). Anyway, this is a *complex-valued* function (OK, it would be possible to consider its real and imaginary parts as a pair, but it is just too cumbersome):

$$C_x(t) = E\left[e^{itX}\right] = E\left[\cos(tX)\right] + iE\left[\sin(tX)\right]$$

Clearly, using power expansion again, it is easy to see that this function also characterizes the distribution of X. Its value lies in the fact that it can be defined for *all* R.V.'s, while the Moment Generating Function requires some restrictions, when we operate in full generality.

Again, in full generality, recovering a distribution from its characteristic function is equivalent to inverting a Fourier transform, which is an advanced problem.

2.3.3 Generating Functions

There is a special group of discrete R.V.'s for which the characteristic function can be rewritten in a suggestive way. Namely, consider a R.V. that takes only non negative integers as values. To cover all cases, we set

$$p_i = P[X = i] \qquad i = 0, 1, 2, \ldots$$

for all integers. We can include the case of integer-valued variables that take on infinitely many values. In any case, with some of the p_i possibly being zero, we will have

$$\sum_{j=0}^{\infty} p_j = 1$$

Let us write now the CF of such a R.V.:

$$E\left[e^{itX}\right] = E\left[\left(e^{it}\right)^X\right] = E\left[z^X\right] = H_X(z) = \sum_{j=0}^{\infty} p_j z^j$$

if we define the complex number $z = e^{it} = \cos t + i \sin t$. A direct check shows that $|z| = \cos^2 t + \sin^2 t = 1$. We can actually let z take values such that $|z| < 1$ too, since H_X is defined as a power series whose radius of convergence is no less than 1, since setting $z = 1$ yields $H_X(1) = \sum_{j=0}^{\infty} p_j = 1$.

The interesting thing about the generating function is that the coefficients of its power expansion yield the distribution weights directly.

3 Calculating Distributions

The point of introducing tools like those in sec. 2.3 is that they allow (sometimes) to get the distribution of a R.V. relatively easy.

You will have noticed that the functions we have used are all of the form e^{aX} where a is a real or complex number. Suppose that (as often is the case) we are interested in calculating the distribution of $X + Y$, where X and Y are random variables. We can try to go through one of these generating functions, and use the main property of exponentials:

$$E\left[e^{a(X+Y)}\right] = E\left[e^{aX}e^{aY}\right] = \sum_{x,y} P[X = x, Y = y]\, e^{ax}e^{ay}$$

Now, this may not be much easier than through some other route, but, *in the very special case when X and Y are independent,*

$$P[X = x, Y = y] = P[X = x]\, P[Y = y]$$

(by definition), and so

$$M_{X+Y} = M_X M_Y$$

where M is any of the generating functions introduced in sec. 2.3. Please, note that this handy fact **only holds in the case of independent R.V.'s!**

3.1 An example: Summing Independent Bernoulli Variables

Consider any number of *independent* R.V. X_i, all with the distribution

$$P[X_i = 1] = p; P[X_i = 0] = 1 - p$$

Variables with this distribution are called "Bernoulli R.V.'s". It is easy to compute all their functions (these only depend on the distribution, so they are all equal):

$$M_{X_1}(t) = pe^{1\cdot t} + (1-p)\, e^{0\cdot t} = 1 + p\left(e^t - 1\right)$$

$$C_{X_1}(t) = pe^{it\cdot 1} + (1-p)\, e^{it\cdot 0} = 1 + p\left(e^{it} - 1\right)$$

$$H_{X_1}(z) = (1-p) + pz = 1 + p(z-1)$$

Now, if we add n of these independent variables, we will have (looking, for simplicity, at the generating function only: the others are just similar):

$$H_{\sum_1^n X_j}(z) = (H_{X_1})^n = ((1-p)+pz)^n = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) z^k p^k (1-p)^{n-k}$$

Recalling how we reconstruct the distribution from the generating function, we have

$$P\left[\sum_{j=1}^{n} X_j = k\right] = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}$$

This distribution is called the "Binomial Distribution" with parameters n, p, or $b(n,p)$. For instance, if you are playing n independent games with constant probability p of winning, this is he distribution of the numbers of wins you score.

As an exercise (but it is obvious from the construction), you can check that the sum of two independent R.V.'s with distributions $b(n,p)$, and $b(m,p)$ has distribution $b(n+m,p)$.

3.2 Variation: "Spin" Variables

A variation on the Bernoulli distribution (where R.V.'s take the values 0 and 1) is the distribution

$$P[X_j = 1] = p; P[X_j = -1] = 1 - p$$

This is sometimes preferred in physics, where (in Quantum Mechanics) such variables appear naturally. As above, we calculate

$$M_{X_1}(t) = pe^t + (1-p)\, e^{-t} = e^{-t} + p\left(e^t - e^{-t}\right) = e^{-t} + 2p\sinh t$$

$$C_{X_1}(t) = pe^{it} + (1-p)\, e^{-it} = e^{-it} + p\left(e^{it} - e^{-it}\right) = e^{-t} + 2ip\sin t$$

$$H_{X_1}(z) = (1-p)\, z^{-1} + pz$$

Since our variables do not take only non negative integer values, we find ourselves with a *Laurent* series and a pole at $z = 0$ for our generating function. Suspending disbelief about the mathematical solidity of our procedure until you go through a class in Complex Analysis, we proceed to find the distribution of the sum

of n independent such variables just as before: the corresponding generating function will be

$$\left((1-p)\,z^{-1}+pz\right)^n=\sum_{k=0}^{n}\left(\begin{array}{c} n \\ k \end{array}\right)z^k z^{k-n}p^k\,(1-p)^{n-k}=$$

$$=\sum_{j=-n}^{j=n}z^j\left(\begin{array}{c} n \\ \frac{n+j}{2} \end{array}\right)p^{\frac{n+j}{2}}\,(1-p)^{\frac{n-j}{2}}$$

To read the formula we have to remember that if n is even, the sum can only take even values, so both n and j will be even only and $\frac{n\pm j}{2}$ are integers, while if n is odd, the sum, hence j, only takes odd values, and, again $\frac{n\pm j}{2}$ are integers. In other words, we read $\left(\begin{array}{c} n \\ \frac{n+j}{2} \end{array}\right)=0$ whenever $\frac{n+j}{2}$ is not an integer.

Week 5 Errata

Page 2, Sec., 2.1.1

A misprint in line 11 should be obvious. It should read

"Clearly, if X takes on the values $x_1 < x_2 < \ldots < x_n$, and $x_k \leq x_{k+1}$"

Page 3/4 "The Moment Problem"

It could have been written better. In case it seems unclear, the statement means that the "moment problem" (reconstructing the distribution of a random variable from the sole knowledge of its moments) is solvable for discrete finite random variables (as explained there, it amounts to solving a linear system), but is much more difficult, and possibly unsolvable (i.e., the moments may not determine the distribution uniquely) in more general situations. As a matter of fact, in more general situations, moments may even not be defined at all, as we will see in the next chapter. Even when they are, though, some conditions need to be satisfied for them to identify the distribution (for example, one such condition is that the moment generating function exist and be real analytic).