Lecture 5: Statistical Independence, Discrete Random Variables

14 September 2005

1 Statistical Independence

If

$$\Pr(A|B) = \Pr(A)$$

we say that A is *statistically independent* of B: whether B happens makes no difference to how often A happens

Since $\Pr(A \cap B) = \Pr(A|B)\Pr(B)$, if A is independent of B, then

$$\Pr(A \cap B) = \Pr(A)\Pr(B)$$

If this holds, though, then B is also independent of A:

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)} = \frac{\Pr(A)\Pr(B)}{\Pr(A)} = \Pr(B)$$

so we can just say "A and B are independent".

Note that events can be logically or physically independent but still statistically dependent. Let $A =$ "scored above 700 on the math SAT" and $B =$ "attends CMU". These are logically independent (neither one implies the other), but statistically quite dependent, because $\Pr(A|B) > \Pr(A)$.

Statistical independence means one event conveys no information about the other; statistical dependence means there is some information. Making this precise is the subject of information theory. Information theory is my area of research, so if I start talking about it I won't shut up; so I won't start.

Statistically independent is *not* the same as mutually exclusive: if A and B are mutually exclusive, then they can't be independent, unless one of them is probability 0 to start with:

$$\Pr(() A \cap B) = 0 = \Pr(() A)\Pr(() B)$$

"Mutually exclusive" is definitely informative: if one happens, then the other can't. (They could still both not happen, unless they're jointly exhaustive.)

2 Discrete Random Variables

A **random variable** is just one which is the result of a random process, like an experimental measurement subject to noise, or a reliable measurement of some fluctuating quantity. What this means in practice is that there is a certain probability for the random variable X to take on any particular value in the sample space.

By convention, we'll use capital letters, $X, Y, Z, W, \ldots$ for random variables, and the corresponding lower-case letters for points in the sample space — particular outcomes or **realizations** of the random variable. The difference between X and x is the difference between "the sum of two dice" and "5".

If the sample space is discrete, we completely specify the random variable by giving the probability for each elementary outcome or realization, $\Pr(X = x)$. This is often abbreviated $p(x)$, and called the **probability distribution** or **probability distribution function**. Because it's a probability, $p(x) \geq 0$ for all x, and $\sum_x p(x) = 1$. Conversely, any function which satisfies those two rules can be a probability distribution.

The probability distribution is also sometimes called the **probability mass function** (p.m.f.), on the analogy of having a unit mass to spread over the sample space.

Any function of a random variable is again a random variable. But it may be a trivial one, like $\sin^2 X + \cos^2 X$.

2.1 Expectation

The expectation of a random variable is its average value, with weights in the average given by the probability distribution

$$\mathbf{E}[X] \equiv \sum_x \Pr(X = x)\, x$$

Continuing the mass analogy from the probability mass function, the expectation is the location of the center of mass.

Expectation is like the population mean, so the basic properties of the mean carry over:
$\mathbf{E}[aX + b] = a\mathbf{E}[X] + b$
If $X \geq Y$, then $\mathbf{E}[X] \geq \mathbf{E}[Y]$
If we want to know the expectation of a function of X, $\mathbf{E}[f(X)]$, we just apply the formula:

$$\mathbf{E}[f(X)] = \sum_x \Pr(X = x)\, f(x)$$

The variance is the expectation of $(X - \mathbf{E}[X])^2$.

$$\mathrm{Var}(X)] \equiv \sum_x p(x)(x - \mathbf{E}[X])^2$$

$\mathrm{Var}(X)$ gives an indication of just how much spread there is in the population around the average (expectation) value — how much slop we should anticipate around the expectation.

(The variance is the moment of inertia around the center of mass.)

3 Bernoulli Random Variables

A **Bernoulli** random variable is just one which takes on the values 0 or 1. We completely specify the distribution with one number, $\Pr(X = 1) = p$. Bernoulli variables are really simple, but a lot of more interesting things can be represented using them, so they're an important place to start.

The expectation is p, and the variance is $p(1-p)$. Let's see that in detail:

$$\begin{aligned}
\mathbf{E}[X] &= \sum_{x=0}^{1} xp(x) \\
&= 0 \times p(0) + 1 \times p(1) \\
&= p \\
\mathrm{Var}(X) &= \mathbf{E}\left[X^2\right] - (\mathbf{E}[X])^2 \\
&= \left(0^2 p(0) + 1^2 p(1)\right) - (p)^2 \\
&= p - p^2 \\
&= p(1-p)
\end{aligned}$$

Notice that if we change p, we get a different probability distribution, but always one which has the same mathematical form. Also, the mean, variance, and all the other properties of X depend only on p. We say that p is the **parameter** of the Bernoulli distribution.

3.1 Indicators

For any event A, we can define a Bernoulli variable which is 1 when A happens and 0 otherwise. This is the indicator variable or indicator function for A, 1_A. So $\Pr(A) = \mathbf{E}[1_A]$. These prove useful, because in some situations it's easier for us to calculation $\mathbf{E}[1_A]$, than to directly find $\Pr(A)$.

4 Sums of Random Variables

Suppose we have two random variables, which we'll call X and Y. We can add them, to get a new variable, $X + Y$. It will also be random. It will have some expectation, $\mathbf{E}[X + Y]$. What is that?

$$\mathbf{E}[X+Y] = \sum_{x,y} (x+y)\Pr(X = x, Y = y)$$

$$
\begin{aligned}
&= \sum_{x,y} x\Pr(X=x, Y=y) + \sum_{x,y} y\Pr(X=x, Y=y) \\
&= \sum_{x} x \sum_{y} \Pr(X=x, Y=y) + \sum_{y} y \sum_{x} \Pr(X=x, Y=y)
\end{aligned}
$$

By total probability, $\sum_y \Pr(X=x, Y=y) = \Pr(X=x)$, likewise $\sum_x \Pr(X=x, Y=y) = \Pr(Y=y)$. So,

$$
\begin{aligned}
\mathbf{E}[X+Y] &= \sum_{x} x\Pr(X=x) + \sum_{y} y\Pr(Y=y) \\
&= \mathbf{E}[X] + \mathbf{E}[Y]
\end{aligned}
$$

5 Binomial

Finally, a use for all that combinatorics!

Imagine we have n "trials" or units, each of which can succeed (or be 1) with probability p — that is, each trial is a Bernoulli variable. Assume the trials are independent. Our random variable X is the number of successes, or the sum of the Bernoulli variables. What is the distribution of X?

Let's start by thinking of a simple example, where $n = 3$. If $X = 0$, it must be the case that none of the three trials succeeded. There is only one sequence of outcomes which will do this: $(0, 0, 0)$. The probability of this event is thus $(1-p)^3$. Now consider $X = 1$. We can get this by the outcomes $(0, 0, 1)$, $(0, 1, 0)$, and $(1, 0, 0)$. All three are equally likely: they have probability $p(1-p)^2$, since there's one success (at probability p) and two failures (at $1-p$ each). $X = 2$, similarly, has three possibilities, each of which has probability $p^2(1-p)$, and $X = 3$ only one, probability p^3. Clearly $X < 0$ and $X > 3$ are both impossible. So we have $\Pr(X=x) = \binom{3}{x} p^x (1-p)^{3-x}$, if $0 \leq x \leq n$, and $\Pr(X=x) = 0$ otherwise.

This is going to be generally true, whatever n is: the probability that $X = x$ will be the number of n trials sequences with x success, times the probability of any one such sequence, or

$$
\Pr(X=x) = \binom{n}{x} p^x (1-p)^{n-k}
$$

Notice that this has *two* parameters, n and p.

Because a binomial variable with parameters n, p is the sum of n independent random variables with parameter p, we can find the expectation very simply: $\mathbf{E}[X] = np$. The alternative to this is a quite ugly sum.