Lecture 6: Discrete Random Variables

19 September 2005

1 Expectation

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

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

If c is a constant, $\mathbf{E}[c] = c$.
If a and b are constants, $\mathbf{E}[aX + b] = a\mathbf{E}[X] + b$.
If $X \geq Y$, then $\mathbf{E}[X] \geq \mathbf{E}[Y]$
Now let's think about $\mathbf{E}[X + Y]$.

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

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

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

Notice that $\mathbf{E}[X]$ works just like a mean; in fact we can think of it as being the population mean (as opposed to the sample mean).

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

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

which we can show is $\mathbf{E}\left[X^2\right] - (\mathbf{E}[X])^2$.

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbf{E}\left[(X - \mathbf{E}[X])^2\right] \\
&= \mathbf{E}\left[X^2 - 2X\mathbf{E}[X] + (\mathbf{E}[X])^2\right] \\
&= \mathbf{E}\left[X^2\right] - \mathbf{E}[2X\mathbf{E}[X]] + \mathbf{E}\left[(\mathbf{E}[X])^2\right]
\end{aligned}$$

Now $\mathbf{E}[X]$ is just another constant, so $\mathbf{E}\left[(\mathbf{E}[X])^2\right] = (\mathbf{E}[X])^2$, and $\mathbf{E}[2X\mathbf{E}[X]] = 2\mathbf{E}[X]\,\mathbf{E}[X] = 2(\mathbf{E}[X])^2$. So

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbf{E}\left[X^2\right] - 2(\mathbf{E}[X])^2 + (\mathbf{E}[X])^2 \\
&= \mathbf{E}\left[X^2\right] - (\mathbf{E}[X])^2
\end{aligned}$$

as promised.

The main rule for variance is this:

$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$$

It's not generally true that $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$; we'll see when it's true later.

1.1 Some useful results

A basic result about expectations is the **Markov inequality**: if X is a non-negative random variable, and a is a positive constant, then

$$\Pr(X \geq a) \leq \frac{\mathbf{E}[X]}{a}$$

Proof: Let $A = \{X \geq a\}$. So $X \geq a1_A$: either $1_A = 0$, in which case $X \geq 0$, or else $1_A = 1$, but then $X \geq a$. So $\mathbf{E}[X] \geq \mathbf{E}[a1_A] = a\mathbf{E}[1_A] = a\Pr(X \geq a)$.

The **Chebyshev inequality** is a special case of the Markov inequality, but a very useful one. It's plain that $(X - \mathbf{E}[X])^2 \geq 0$, so applying the Markov inequality gives

$$\Pr\left((X - \mathbf{E}[X])^2 \geq a^2\right) \leq \frac{\mathrm{Var}(X)}{a^2}$$

Taking the square root of the term inside the left-hand side,

$$\Pr(|X - \mathbf{E}[X]| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}$$

The Chebyshev inequality helps give meaning to the variance: it tells us about how unlikely it is for the random variable to depart very far from its mean.

2 Independent r.v.s

We'll say that random variables are independent if their probability distributions factor, $\Pr(X = x, Y = y) = \Pr(X = x)\Pr(Y = y)$.

If the variables are independent, then $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$.

$$\begin{aligned}
\mathbf{E}[XY] &= \sum_{x,y} xy\Pr(X = x, Y = y) \\
&= \sum_x \sum_y xy\Pr(X = x, Y = y) \\
&= \sum_x \sum_y xy\Pr(X = x)\Pr(Y = y) \\
&= \sum_x x\Pr(X = x) \sum_y y\Pr(Y = y) \\
&= \sum_x x\Pr(X = x)\mathbf{E}[Y] \\
&= \mathbf{E}[Y] \sum_x x\Pr(X = x) \\
&= \mathbf{E}[Y]\mathbf{E}[X]
\end{aligned}$$

This isn't the only time that $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$, though.

Here's where independence gets important: what's the variance of $X + Y$?

$$\begin{aligned}
\mathrm{Var}(X + Y) &= \mathbf{E}\left[(X + Y)^2\right] - (\mathbf{E}[X + Y])^2 \\
&= \mathbf{E}\left[X^2 + 2XY + Y^2\right] - (\mathbf{E}[X] + \mathbf{E}[Y])^2 \\
&= \mathbf{E}\left[X^2\right] + 2\mathbf{E}[XY] + \mathbf{E}\left[Y^2\right] - \left[(\mathbf{E}[X])^2 + 2\mathbf{E}[X]\mathbf{E}[Y] + (\mathbf{E}[Y])^2\right] \\
&= \mathbf{E}\left[X^2\right] - (\mathbf{E}[X])^2 + \mathbf{E}\left[Y^2\right] - (\mathbf{E}[Y])^2 + 2\mathbf{E}[XY] - 2\mathbf{E}[X]\mathbf{E}[Y] \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathbf{E}[XY] - 2\mathbf{E}[X]\mathbf{E}[Y]
\end{aligned}$$

But we've just seen that $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$ if X and Y are independent, so then

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

3 Binomial random variables

Recall that the distribution of the binomial is

$$ProbX = x = \binom{n}{x} p^x (1 - p)^{n-x}$$

and that it's the sum of n independent Bernoulli variables with parameter p.

How do we know this is a valid probability distribution? It's clearly ≥ 0 for all x, but how do I know it sums to 1? Because of the binomial theorem from algebra (which is where the name comes from).

$$\begin{aligned}
(a+b)^n &= \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} \\
(p+(1-p))^n &= \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \\
1^n &= \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \\
1 &= \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k}
\end{aligned}$$

To find the mean and variance, we could either do the appropriate sums explicitly, which means using ugly tricks about the binomial formula; or we could use the fact that X is a sum of n independent Bernoulli variables. Because the Bernoulli variables have expectation p, $\mathbf{E}[X] = np$. Because they have variance $p(1-p)$, $\mathrm{Var}(X) = np(1-p)$.

4 Geometric random variables

Suppose we keep trying independent Bernoulli variables until we have a success; each has probability of success p. Then the probability that the number of failures is k is $(1-p)^k p$. (Be careful, some people use p as the probability of failure here, i.e. they reverse p and $1-p$.)

First, check that this weird thing is a valid probability distribution — does it sum to one? Yes:

$$\sum_{k=0}^{\infty} (1-p)^k p = p \sum_{k=0}^{\infty} (1-p)^k = p \frac{1}{1-(1-p)} = p\frac{1}{p} = 1$$

This uses the geometric series, the fact that $\sum p^k = 1/(1-p)$, if p is between 0 and 1.

Now let's think about the mean.

$$\mathbf{E}[X] = \sum_{k=0}^{\infty} k(1-p)^k p = p \sum_{k=1}^{\infty} k(1-p)^k = p\frac{1}{p^2} = \frac{1}{p}$$

Similarly (but more involvedly) the variance is $(1-p)/p^2$.

4.1 Negative binomial random variables

Instead of just getting one success, we might keep going until we get r of them. The probability distribution then is just $\Pr(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}$, $k \geq r$.

If we think of W_1 as the number of trials we have to make to get the first success, and then W_2 the number of further trials to the second success, and so on, we can see that $X = W_1 + W_2 + \ldots + W_r$, and that the W_i are independent and geometric random variables. So $\mathbf{E}[X] = r/p$, and $\mathrm{Var}(X) = r(1-p)/p^2$.

5 Poisson random variables

Think about a very large number of Bernoulli trials, where $n \to \infty$, but the expected number of successes stays constant, say λ. For instance, suppose we're looking at the number of particles emitted by a chunk of radioactive substance over one-second intervals of time. Every atom has a certain probability to decay over a given unit of time; as we make the time intervals smaller, we make those probabilities smaller, but the average total should still come to the same number.

If we have only a finite n, but n is very large, so $p = \lambda/n$.

$$\begin{aligned}
\Pr(X = k) &= \binom{n}{k} p^k (1-p)^k \\
&= \frac{n!}{k!(n-k)!} \frac{p^k}{(1-p)^k} (1-p)^n
\end{aligned}$$

Since n is large, we can use Stirling's approximation on $n!$ and $(n-k)!$, so $n! \approx n^n$ and $(n-k)! \approx (n-k)^{n-k} \approx n^{n-k}$.

$$\begin{aligned}
\Pr(X = k) &\approx \frac{n^k}{k!} \frac{p^k}{(1-p)^k} \left(1 - \frac{\lambda}{n}\right)^n \\
&\to \frac{\lambda^k}{k!} e^{-\lambda}
\end{aligned}$$

because $\lim (1 + x/n)^n = e^x$.

We can check that the probability adds up to one, because

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}$$

We can also get the mean:

$$\begin{aligned}
\mathbf{E}[X] &= \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} \\
&= \sum_{k=1}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} \\
&= \sum_{k=1}^{\infty} \lambda \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda}
\end{aligned}$$

$$
\begin{aligned}
&= \lambda \sum_{k=1}^{\infty} \lambda \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} \\
&= \lambda \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} \\
&= \lambda
\end{aligned}
$$

The easiest way to get the variance is to first calculate $\mathbf{E}[X(X-1)]$, because this will let us use the same sort of trick about factorials and the exponential function again.

$$
\begin{aligned}
\mathbf{E}[X(X-1)] &= \sum_{k=0}^{\infty} k(k-1) \frac{\lambda^k}{k!} e^{-\lambda} \\
\mathbf{E}\left[X^2 - X\right] &= \sum_{k=2}^{\infty} \frac{\lambda^k}{(k-2)!} e^{-\lambda} \\
\mathbf{E}\left[X^2\right] - \mathbf{E}[X] &= \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} e^{-\lambda} \\
&= \lambda^2 \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} \\
&= \lambda^2
\end{aligned}
$$

So $\mathbf{E}\left[X^2\right] = \mathbf{E}[X] + \lambda^2$.

$$
\begin{aligned}
\mathrm{Var}(X) &= \mathbf{E}\left[X^2\right] - (\mathbf{E}[X])^2 \\
&= \mathbf{E}[X] + \lambda^2 - \lambda^2 \\
&= \mathbf{E}[X] = \lambda
\end{aligned}
$$

6 Adding Many Independent Random Variables

Remember the Chebyshev inequality:

$$
\Pr(|X - \mathbf{E}[X]| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}
$$

Let's look at the sum of a whole bunch of independent random variables with the same distribution, $S_n = \sum_{i=1}^{n} X_i$.

We know that $\mathbf{E}[S_n] = \mathbf{E}[\sum_{i=1}^{n} X_i] = \sum \mathbf{E}[X_i] = n\mathbf{E}[X_1]$, because they all have the same expectation. Because they're independent, and all have the same variance, $\mathrm{Var}(S_n) = n\mathrm{Var}(X_1)$. So

$$
\Pr(|S_n - n\mathbf{E}[X_1]| \geq a) \leq \frac{n\mathrm{Var}(X_1)}{a^2}
$$

Now, notice that S_n/n is just the sample mean, if we draw a sample of size n. So we can use the Chebyshev inequality to estimate the chance that the sample mean is far from the true, population mean, which is the expectation.

$$\begin{aligned}
\Pr\left(\left|\frac{S_n}{n} - \mathbf{E}\left[X_1\right]\right| \geq \epsilon\right) &= \Pr\left(\left|S_n - n\mathbf{E}\left[X_1\right]\right| \geq n\epsilon\right) \\
&\leq \frac{n\mathrm{Var}\left(X_1\right)}{n^2\epsilon^2} \\
&\leq \frac{\mathrm{Var}\left(X_1\right)}{n\epsilon^2}
\end{aligned}$$

Observe that whatever ϵ is, the probability must go to zero like $1/n$ (or faster). So the probability that the sample mean differs from the population mean by as much as ϵ can be made arbitrarily small, by taking a large enough sample. This is the **law of large numbers**.