EE 178/278A Probabilistic Systems Analysis

Spring 2014 Tse/Hussami Lecture 7

In the last lecture, we covered random variables and probability mass functions. The probability mass function (or distribution) assigns a mass to each point which is equal to the probability of that point. The distribution of a random variable summarizes all the probabilistic information about it. However, the complete distribution can be complicated or very hard to calculate. Moreover, even when we can compute the complete distribution of a r.v., it's not always very informative. Instead, we would like to have a few simple numbers which summarize the main features of the distribution.

For example, suppose that we have the histogram of all the student scores on an exam. We might be more interested in knowing the average score, and the variation around that average. We will now introduce two important concepts that capture the main properties of the distribution of a random variable.

Expectation

For the reasons mentioned above, we seek to *compress* the distribution into a more compact, convenient form that is also easier to compute. The most widely used such form is the *expectation* (or *mean*) of the r.v.

Definition 7.1 (expectation): The *expectation or mean* of a discrete random variable X is defined as

$$\mathbb{E}[X] = \sum_{a \in \mathscr{A}} a \times \mathbf{P}(X = a),$$

where the sum is over all possible values taken by the r.v.

The expectation can be interpreted as being the center of mass. It can be seen in some sense as a "typical" value of the r.v. (though note that it may not actually be a value that the r.v. ever takes on). The question of how typical the expectation is for a given r.v. is a very important one that we shall return to in a later lecture.

Other definitions can be used to capture features of the random variable. For example, we could be interested in the mode of the distribution, which is the value that achieves the highest probability. But it turns out that the mean has certain properties that make it the most important measure.

An alternative way of defining expectation is

$$\mathbb{E}[X] = \sum_{\omega} X(\omega)\mathbf{P}(\omega).$$

The two definitions are equivalent. This definition looks at the individual outcomes in each of the events. The difference is, that in the first definition, the probability of each event $\mathbf{P}(X = a)$ is computed beforehand which is the sum of $\mathbf{P}(\omega)$ over those sample points ω for which $X(\omega) = a$. And we know that every sample point $\omega \in \Omega$ is in exactly one of these events $X = a$.

To understand the difference in computing the expectation using the two definitions, we can again consider a histogram showing the scores students obtained on an exam. Using the first definition, the expectation is computed by weighting each score by the fraction of students that got it. However, if we use the alternative definition, we add up the score of each student separately.

Here are some simple examples of expectations.

1. **Single die.** Throw one fair die. Let X be the number that comes up. Then X takes on values $1,2,\dots,6$ each with probability $\frac{1}{6}$. More explicitly, this random variable is defined as $X(\omega)=\omega$.

   $$\mathbb{E}[X]=\frac{1}{6}(1+2+3+4+5+6)=\frac{21}{6}=\frac{7}{2}.$$

   Note that X never actually takes on its expected value $\frac{7}{2}$. (Because each event $X=a$ has exactly one outcome in it, there is no difference in using the two definitions in computing the expectation.)

2. **Two dice.** Throw two fair dice. Let X be the sum of their scores. Therefore X is defined as $X((\omega_1,\omega_2))=\omega_1+\omega_2$.There are 36 possible pairs of scores. The distribution of X is

a	2	3	4	5	6	7	8	9	10	11	12
$\mathbf{P}(X=a)$	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$

   The expectation is therefore

   $$\mathbb{E}[X]=\left(2\times\frac{1}{36}\right)+\left(3\times\frac{1}{18}\right)+\left(4\times\frac{1}{12}\right)+\cdots+\left(12\times\frac{1}{36}\right)=7. \tag{1}$$

   The expectation can also be computed using the alternative definition in the following way:

   $$\mathbb{E}[X]=2\times\frac{1}{36}+3\times\frac{1}{36}+3\times\frac{1}{36}+4\times\frac{1}{36}+\dots \tag{2}$$

   where we consider each outcome separately. Note that the second and third term (**??**) correspond to the second term in (1).

Now let us consider a few general questions regarding distributions, random variables and expectation. Does a probability have a unit? No, it is a normalized number between 0 and 1. Does a random variable have a unit? Yes, it might. For example the random variable can be measuring the height of individuals. In that case, the individuals are the outcomes, and the random variable could be in meters. Does the expectation of X have a unit? Yes, it has the same unit as the random variable.

We will now turn to useful properties of expectation.

Linearity of expectation

So far, we've computed expectations by brute force: i.e., we have written down the whole distribution and then added up the contributions for all possible values of the r.v.,, or enumerate all possible outcomes and added up all the contributions from each outcome. The real power of expectations is that in many real-life examples they can be computed much more easily using a simple shortcut. The shortcut is the following:

Theorem 7.1: *For any two random variables X and Y on the same probability space, we have*

$$\mathbb{E}[X+Y]=\mathbb{E}[X]+\mathbb{E}[Y].$$

Also, for any constant c, we have

$$\mathbb{E}[cX] = c\mathbb{E}[X].$$

A note: here $X+Y$ denotes the random variable that is the sum of X and Y, i.e., it takes on the value $X(\omega)+Y(\omega)$ at the outcome ω, cX denotes another random variables that is a scaled version of X.

Proof: Let's write out $\mathbb{E}[X+Y]$ using the alternative definition of expectation:

$$\begin{aligned}\mathbb{E}[X+Y] &= \sum_{\omega\in\Omega}(X+Y)(\omega)\times\mathbf{P}[\omega]\\ &= \sum_{\omega\in\Omega}(X(\omega)+Y(\omega))\times\mathbf{P}[\omega]\\ &= \sum_{\omega\in\Omega}\Big(X(\omega)\times\mathbf{P}[\omega]\Big)+\sum_{\omega\in\Omega}\Big(Y(\omega)\times\mathbf{P}[\omega]\Big)\\ &= \mathbb{E}[X]+\mathbb{E}[Y].\end{aligned}$$

This completes the proof of the first equality. The proof of the second equality is left as an exercise. □

Theorem 7.1 is very powerful: it says that the expectation of a sum of r.v.'s is the sum of their expectations, no matter what those r.v.'s may be. We can use Theorem 7.1 to conclude things like $\mathbb{E}[3X-5Y] = 3\mathbb{E}[X]-5\mathbb{E}[Y]$. This property is known as *linearity of expectation*. One important caveat: Theorem 7.1 does *not* say that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, or that $\mathbb{E}[\frac{1}{X}] = \frac{1}{\mathbb{E}[X]}$ etc. These claims are not true in general. It is only sums and differences and constant multiples of random variables that behave so nicely.

Now let's see some examples of Theorem 7.1 in action.

1. **Two dice again.** Here's a much less painful way of computing $\mathbb{E}[X]$, where X is the sum of the scores of the two dice. Note that $X = X_1+X_2$, where X_i is the score on die i. We know from example 1 above that $\mathbb{E}[X_1] = \mathbb{E}[X_2] = \frac{7}{2}$. So by Theorem 7.1 we have $\mathbb{E}[X] = \mathbb{E}[X_1]+\mathbb{E}[X_2] = 7$.

2. Let's go back and answer our original question about the class of n students. Overall, there are $n!$ equally likely permutations of the homeworks. Recall that the random variable X is the number of students who receive their own homework after shuffling (or equivalently, the number of fixed points). A natural question to ask if we do this procedure is: what is the average number of students that pick up the correct homework? We want to compute the expectation of X to summarize the performance of the scheme. $n!$ is a large number so we do not want to compute the average by counting the number of students that get their homework back correctly for every possible outcome.

   The total number of outcomes in the sample space expands with n, however, the number of outcomes in which we have students that get their own homework back is also increasing. It is not clear that as the sample space expands, the expected value of X is lower. The chance of a single student getting his homework back is $\frac{1}{n}$, which is decreasing in n. This seems to suggest the expected value will decrease. But, we are not looking at whether a specific student gets his homework back. We are interested in the expected number of students that get it back.

   By using linearity, we can compute this expectation very easily. To take advantage of Theorem 7.1, we need to write X as the *sum* of simpler r.v.'s. But since X *counts* the number of times something happens, we can write it as a sum using the following trick:

$$X = X_1+X_2+\cdots+X_n, \qquad \text{where } X_i = \begin{cases}1 & \text{if student } i \text{ gets her own homework;}\\ 0 & \text{otherwise}\end{cases} \tag{3}$$

[You should think about this equation for a moment. Remember that all the X's are random variables. What does an equation involving random variables mean? What we mean is that, *at every sample point* ω, we have $X(\omega) = X_1(\omega) + X_2(\omega) + \cdots + X_{20}(\omega)$. Do you see why this is true?] A 0/1-valued random variable such as X_i is called an *indicator* random variable of the corresponding event (in this case, the event that student i gets her own homework). For indicator r.v.'s, the expectation is particularly easy to calculate. Namely,

$$\mathbb{E}[X_i] = (0 \times \mathbf{P}(X_i = 0)) + (1 \times \mathbf{P}(X_i = 1)) = \mathbf{P}(X_i = 1).$$

But in our case, we have

$$\mathbf{P}(X_i = 1) = \mathbf{P}(\text{student } i \text{ gets her own homework}) = \frac{1}{n}.$$

Now we can apply Theorem 7.1 to (3), to get

$$\mathbb{E}[X] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n] = n \times \frac{1}{n} = 1.$$

So we see that the expected number of students who get their own homeworks in a class of size n is 1. Therefore, as you scale the system, the expectation remains the same no matter how large n is. The individual effect and the aggregation effect cancel each other out, and the final result is independent of the system size. *The expected number of fixed points in a random permutation of n items is always 1*, regardless of n. Amazing, but true.

3. Let X be the number of Heads you get from n flips. We have seen that $X \sim Bin(n, p)$:

$$\mathbf{P}(X = i) = \binom{n}{i} p^i (1-p)^i$$

where i goes from 0 to n. Following the definition, the expectation of X can be written as

$$\mathbb{E}[X] = \sum_i i \times p_X(i) = \sum_{i=0}^{n} \binom{n}{i} i p^i (1-p)^{n-i}.$$

However, there is again an easier way to compute this expectation. As in the previous example, we write X as a sum of n random variables, where $X_i = 1$ if *ith* flip is heads, and 0 otherwise. Then, we obtain that $\mathbb{E}[X] = n\mathbb{E}[X_1] = np$.

Variance

The expectation is like the center of the distribution, it gives a rough idea of where the values are. But we are still missing something important. The mean does not describe the randomness of the phenomenon, meaning the variation around the expectation of the random variable. Consider again the scores on an exam and suppose the mean is given to be 50: all the students could have gotten exactly 50, but the scores can also vary from 0 to 100 as long as they still satisfy the mean constraint. The spread of the random variable around its mean is related to how much uncertainty there is in the random variable. We will now define a notion of spread of the random variable.

We are interested in summarizing the variation of the random variable around its mean. Let $\mu = \mathbb{E}[X]$. One possible measure of deviation from the mean of X is $X - \mu$.. The deviation can be small or large depending on the values the random variable takes on. So $X - \mu$ is itself a random variable, and we can compute its mean to get the average of the deviation. However, $\mathbb{E}[X-\mu]$ is equal to zero by linearity. This happens because it takes on positive and negative values which cancel out. So actually $X-\mu$ is not a good measure of deviation, Instead, we could square $X-\mu$ or take its absolute value as a better measure of deviation, as then it will be non-negative. Unfortunately, computing the expected value of $|X-\mu|$ turns out to be a little awkward, due to the absolute value operator. Therefore we consider the random variable $(X-\mu)^2$ as a measure of deviation of X around the mean. We take its expectation and get the variance of the random variable to be:

Definition 7.2 (variance): The *variance* of a discrete random variable X is defined as

$$\mathrm{Var}(X) = \mathbb{E}[(X-\mu)^2].$$

If X measures the height in *cm*, the unit of the variance is cm^2. So we also define the standard deviation of a random variable.

Definition 7.3 (Standard deviation): The *standard deviation* of a discrete random variable X is defined as

$$\sigma_X = \sqrt{\mathrm{Var}(X)}$$

which has the same unit as the random variable X.