Conditional Expectations

David S. Rosenberg

Note: Although we usually use lower case variables for both random variables and the values they take, in this note we'll follow the convention common in probability theory that random variables are denoted with capital letters, and the values they take on are lower case.

- Suppose X and Y are random variables taking values in spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively. If X and Y have a joint distribution given by $(X, Y) \sim P_{X,Y}$, how should we understand the expression $\mathbb{E}[Y \mid X]$ or $\mathbb{E}[Y \mid x]$ or $\mathbb{E}[Y \mid X = x]$?

- Let's start with understanding $\mathbb{E}[Y]$. Even though we write expectations in terms of random variables, you can also think of them as properties of distributions. For simplicity, suppose $\mathcal{Y}$ is a discrete subset of $\mathbf{R}$ (such as the integers), and random variable Y has probability mass function (PMF) $p(y)$. Then
$$\mathbb{E}\left[Y\right] = \sum_{y \in \mathcal{Y}} y p(y).$$

- We can interpret $\mathbb{E}\left[Y|X = x\right]$ or $\mathbb{E}[Y \mid x]$ as the expectation of the conditional distribution of Y given $X = x$. If $\mathcal{X}$ is also a discrete space, then we can write $p(x, y)$ as the joint PMF of X and Y, and the conditional PMF of Y given $X = x$ as
$$p(y \mid x) = \frac{p(x, y)}{p(x)}.$$
From basic probability theory, we know that $p(y \mid x)$ is PMF on $\mathcal{Y}$, in the sense that $p(y \mid x) \geq 0 \; \forall y \in \mathcal{Y}$ and $\sum_{y \in \mathcal{Y}} p(y \mid x) = 1$. So $\mathbb{E}\left[Y \mid x\right]$ is the expectation of the distribution corresponding to PMF $p(y \mid x)$:
$$\mathbb{E}\left[Y \mid x\right] = \sum_{y \in \mathcal{Y}} y p(y \mid x).$$

- Note that for every value of x, $p(y \mid x)$ may give a different distribution on $\mathcal{Y}$, and thus they may also have different expectations. So we can view $\mathbb{E}[Y \mid x]$ as a function of x. To emphasize this, let's define the function $f : \mathcal{X} \to \mathbf{R}$ such that $f(x) = \mathbb{E}[Y \mid x]$. Note that there is nothing random about this function. We can plug in any $x \in \mathcal{X}$ to the function, and it always gives us back the same element of $\mathbf{R}$.

- We can now interpret $\mathbb{E}[Y \mid X]$. If f is the function giving $\mathbb{E}[Y \mid x]$, then $\mathbb{E}[Y \mid X] = f(X)$. In other words, it's what we get when we plug in the random variable X to the deterministic function f. Since X is random, $f(X)$ and thus $\mathbb{E}[Y \mid X]$ are themselves random variables. As such, they have distributions. One common thing we often want to do is to find the expectation of this distribution, which we would denote $\mathbb{E}[\mathbb{E}[Y \mid X]]$. This inner expectation is over Y, and the outer expectation is over X. To clarify, this could be written as $\mathbb{E}^X\left[\mathbb{E}^Y[Y \mid X]\right]$, though it is not usually written that way. Note that expression $\mathbb{E}[\mathbb{E}[Y \mid X]]$ is just a number – it's not random, and not a random variable. Let's expand it out a bit:
$$\mathbb{E}[\mathbb{E}[Y \mid X]] = \sum_{x \in \mathcal{X}} p(x)\mathbb{E}[Y \mid X = x]$$

- It turns out there is a "law of iterated expectations" (Wikipedia article) which states that
$$\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[Y].$$
It's easy to prove in the case of finite $\mathcal{X}$ and $\mathcal{Y}$. Continuing the expansion above, we get
$$\begin{aligned}
\mathbb{E}[\mathbb{E}[Y \mid X]] &= \sum_{x \in \mathcal{X}} p(x)\mathbb{E}[Y \mid X = x] \\
&= \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x)y \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} yp(x, y) \\
&= \sum_{y \in \mathcal{Y}} y \sum_{x \in \mathcal{X}} p(x, y) \\
&= \sum_{y \in \mathcal{Y}} yp(y) \\
&= \mathbb{E}Y
\end{aligned}$$