MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.436J/15.085J Fall 2018

Recitation 3

1 CONDITIONAL EXPECTATIONS

We have already defined the notion of a conditional PMF, $p_{X\,|\,Y}(\,\cdot\,|\,y)$, given the value of a random variable Y. Similarly, given an event A, we can define a conditional PMF $p_{X|A}$, by letting $p_{X|A}(x) = \mathbb{P}(X = x\,|\,A)$. In either case, the conditional PMF, as a function of x, is a bona fide PMF (a nonnegative function that sums to one). As such, it is natural to associate a (conditional) expectation to the (conditional) PMF.

Definition 1. *Given an event A, such that $\mathbb{P}(A) > 0$, and a discrete random variable X, the* ***conditional expectation*** *of X given A is defined as*

$$\mathbb{E}[X \mid A] = \sum_x x p_{X\,|\,A}(x),$$

provided that the sum is well-defined.

Note that the preceding also provides a definition for a conditional expectation of the form $\mathbb{E}[X\,|\,Y = y]$, for any y such that $p_Y(y) > 0$: just let A be the event $\{Y = y\}$, which yields

$$\mathbb{E}[X\,|\,Y = y] = \sum_x x p_{X\,|\,Y}(x\,|\,y).$$

We note that the conditional expectation is always well defined when either the random variable X is nonnegative, or when the random variable X is integrable. In particular, whenever $\mathbb{E}[|X|] < \infty$, we also have $\mathbb{E}[|X|\,|\,Y = y] < \infty$, for every y such that $p_Y(y) > 0$. To verify the latter assertion, note that for every y such that $p_Y(y) > 0$, we have

$$\sum_x |x| p_{X|Y}(x\,|\,y) = \sum_x |x| \frac{p_{X,Y}(x, y)}{p_Y(y)} \le \frac{1}{p_Y(y)} \sum_x |x| p_X(x) = \frac{\mathbb{E}[|X|]}{p_Y(y)}.$$

The converse, however, is not true: it is possible that $\mathbb{E}[|X|\,|\,Y = y]$ is finite for every y that has positive probability, while $\mathbb{E}[|X|] = \infty$. This is left as an exercise.

The conditional expectation is essentially the same as an ordinary expectation, except that the original PMF is replaced by the conditional PMF. As such, the conditional expectation inherits all the properties of ordinary expectations (cf. Proposition 4 in the notes for Lecture 6).

1.1 The total expectation theorem

A simple calculation yields

$$\begin{aligned}\sum_y \mathbb{E}[X \mid Y=y] p_Y(y) &= \sum_y \sum_x x p_{X|Y}(x \mid y) p_Y(y) \\ &= \sum_y \sum_x x p_{X,Y}(x,y) \\ &= \mathbb{E}[X].\end{aligned}$$

Note that this calculation is rigorous if X is nonnegative or integrable.

Suppose now that $\{A_i\}$ is a countable family of disjoint events that forms a partition of the probability space Ω. Define a random variable Y by letting $Y = i$ if and only if A_i occurs. Then, $p_Y(i) = \mathbb{P}(A_i)$, and $\mathbb{E}[X \mid Y = i] = \mathbb{E}[X \mid A_i]$, which yields

$$\mathbb{E}[X] = \sum_i \mathbb{E}[X \mid A_i]\mathbb{P}(A_i).$$

Example. (The mean of the geometric.) Let X be a random variable with parameter p, so that $p_X(k) = (1-p)^{k-1}p$, for $p \in \mathbb{N}$. We first observe that the geometric distribution is memoryless: for $k \in \mathbb{N}$, we have

$$\begin{aligned}\mathbb{P}(X-1=k \mid X>1) &= \frac{\mathbb{P}(X=k+1, X>1)}{\mathbb{P}(X>1)} \\ &= \frac{\mathbb{P}(X=k+1)}{\mathbb{P}(X>1)} \\ &= \frac{(1-p)^k p}{1-p} = (1-p)^{k-1}p \\ &= \mathbb{P}(X=k).\end{aligned}$$

In words, in a sequence of repeated i.i.d., trials, given that the first trial was a failure, the distribution of the remaining trials, $X-1$, until the first success is the same as the unconditional distribution of the number of trials, X, until the first success. In particular, $\mathbb{E}[X-1 \mid X>1] = \mathbb{E}[X]$.

Using the total expectation theorem, we can write

$$\mathbb{E}[X] = \mathbb{E}[X \mid X>1]\mathbb{P}(X>1) + \mathbb{E}[X \mid X=1]\mathbb{P}(X=1) = (1+\mathbb{E}[X])(1-p) + 1 \cdot p.$$

We solve for $\mathbb{E}[X]$, and find that $\mathbb{E}[X] = 1/p$.

Similarly,

$$\mathbb{E}[X^2] = \mathbb{E}[X^2 \mid X > 1]\mathbb{P}(X > 1) + \mathbb{E}[X^2 \mid X = 1]\mathbb{P}(X = 1).$$

Note that

$$\mathbb{E}[X^2 \mid X > 1] = \mathbb{E}[(X-1)^2 \mid X > 1] + \mathbb{E}[2(X-1)+1 \mid X > 1] = \mathbb{E}[X^2] + (2/p) + 1.$$

Thus,

$$\mathbb{E}[X^2] = (1-p)(\mathbb{E}[X^2] + (2/p) + 1) + p,$$

which yields

$$\mathbb{E}[X^2] = \frac{2}{p^2} - \frac{1}{p}.$$

We conclude that

$$\operatorname{var}(X) = \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2 = \frac{2}{p^2} - \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$

Example. Suppose we flip a biased coin N times, independently, where N is a Poisson random variable with parameter λ. The probability of heads at each flip is p. Let X be the number of heads, and let Y be the number of tails. Then,

$$\mathbb{E}[X \mid N = n] = \sum_{m=0}^{\infty} m\mathbb{P}(X = m \mid N = n) = \sum_{m=0}^{n} m\binom{n}{m}p^m(1-p)^{n-m}.$$

But X is just the expected number of heads in n trials, so that $\mathbb{E}[X \mid N = n] = np$.

Let us now calculate $\mathbb{E}[N \mid X = m]$. We have

$$\begin{aligned}
\mathbb{E}[N \mid X = m] &= \sum_{n=0}^{\infty} n\mathbb{P}(N = n \mid X = m) \\
&= \sum_{n=m}^{\infty} n\frac{\mathbb{P}(N = n, X = m)}{\mathbb{P}(X = m)} \\
&= \sum_{n=m}^{\infty} n\frac{\mathbb{P}(X = m \mid N = n)\mathbb{P}(N = n)}{\mathbb{P}(X = m)} \\
&= \sum_{n=m}^{\infty} n\frac{\binom{n}{m}p^m(1-p)^{n-m}(\lambda^n/n!)e^{-\lambda}}{\mathbb{P}(X = m)}.
\end{aligned}$$

Recall that $X \overset{d}{=} \text{Pois}(\lambda p)$, so that $\mathbb{P}(X = m) = e^{-\lambda p}(\lambda p)^m/m!$. Thus, after some cancellations, we obtain

$$
\begin{aligned}
\mathbb{E}[N \mid X = m] &= \sum_{n=m}^{\infty} n \frac{(1-p)^{n-m}\lambda^{n-m}e^{-\lambda(1-p)}}{(n-m)!} \\
&= \sum_{n=m}^{\infty} (n-m) \frac{(1-p)^{n-m}\lambda^{n-m}e^{-\lambda(1-p)}}{(n-m)!} \\
&+ m \sum_{n=m}^{\infty} \frac{(1-p)^{n-m}\lambda^{n-m}e^{-\lambda(1-p)}}{(n-m)!} \\
&= \lambda(1-p) + m.
\end{aligned}
$$

A faster way of obtaining this result is as follows. From Theorem 3 in the notes for Lecture 6, we have that X and Y are independent, and that Y is Poisson with parameter $\lambda(1-p)$. Therefore,

$$\mathbb{E}[N \mid X = m] = \mathbb{E}[X \mid X = m] + \mathbb{E}[Y \mid X = m] = m + \mathbb{E}[Y] = m + \lambda(1-p).$$

Exercise. (Simpson's "paradox") Let S be an event and X, Y discrete random variables, all defined on a common probability space. Show that

$$\mathbb{P}[S|X = 0, Y = y] > \mathbb{P}[S|X = 1, Y = y] \qquad \forall y$$

does not imply

$$\mathbb{P}[S|X = 0] \geq \mathbb{P}[S|X = 1]\,.$$

Thus in a clinical trial comparing two treatments (indexed by X) a drug can be more successful on each group of patients (indexed by Y) yet be less successful overall.

1.2 The conditional expectation as a random variable

Let X and Y be two discrete random variables. For any fixed value of y, the expression $\mathbb{E}[X \mid Y = y]$ is a real number, which however depends on y, and can be used to define a function $\phi : \mathbb{R} \to \mathbb{R}$, by letting $\phi(y) = \mathbb{E}[X \mid Y = y]$. Consider now the random variable $\phi(Y)$; this random variable takes the value $\mathbb{E}[X \mid Y = y]$ whenever Y takes the value y, which happens with probability $\mathbb{P}(Y = y)$. This random variable will be denoted as $\mathbb{E}[X \mid Y]$. (Strictly speaking, one needs to verify that this is a measurable function, which is left as an exercise.)

Example. Let us return to the last example and find $\mathbb{E}[X \mid N]$ and $\mathbb{E}[N \mid X]$. We found that $\mathbb{E}[X \mid N = n] = np$. Thus $\mathbb{E}[X \mid N] = Np$, i.e., it is a random variable that takes the value np with probability $\mathbb{P}(N = n) = (\lambda^n/n!)e^{-\lambda}$. We found that $\mathbb{E}[N \mid X = m] = \lambda(1-p) + m$. Thus $\mathbb{E}[N \mid X] = \lambda(1-p) + X$.

Note further that

$$\mathbb{E}[\mathbb{E}[X \mid N]] = \mathbb{E}[Np] = \lambda p = \mathbb{E}[X],$$

and

$$\mathbb{E}[\mathbb{E}[N \mid X]] = \lambda(1-p) + \mathbb{E}[X] = \lambda(1-p) + \lambda p = \lambda = \mathbb{E}[N].$$

This is not a coincidence; the equality $\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$ is always true, as we shall now see. In fact, this is just the total expectation theorem, written in more abstract notation.

Theorem 1. *Let $g : \mathbb{R} \to \mathbb{R}$ be a measurable function such that $Xg(Y)$ is either nonnegative or integrable. Then,*

$$\mathbb{E}\big[\mathbb{E}[X \mid Y]g(Y)\big] = \mathbb{E}[Xg(Y)].$$

In particular, by letting $g(y) = 1$ for all y, we obtain $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$.

Proof: We have

$$\begin{aligned}\mathbb{E}\big[\mathbb{E}[X|Y]g(Y)\big] &= \sum_y \mathbb{E}[X \mid Y = y]g(y)p_Y(y)\\ &= \sum_y \sum_x x p_{X|Y}(x \mid y)g(y)p_Y(y)\\ &= \sum_{x,y} xg(y)p_{X,Y}(x,y) = \mathbb{E}[Xg(Y)].\end{aligned}$$

□

The formula in Theorem 1 can be rewritten in the form

$$\mathbb{E}\big[(\mathbb{E}[X \mid Y] - X)g(Y)\big] = 0. \tag{1}$$

Here is an interpretation. We can think of $\mathbb{E}[X \mid Y]$ as an estimate of X, on the basis of Y, and $\mathbb{E}[X \mid Y] - X$ as an estimation error. The above formula says that the estimation error is uncorrelated with every function of the original data.

Equation (1) can be used as the basis for an abstract definition of conditional expectations. Namely, we define the conditional expectation as a random variable of the form $\phi(Y)$, where ϕ is a measurable function, that has the property

$$\mathbb{E}\big[(\phi(Y) - X)g(Y)\big] = 0,$$

for every measurable function g. The merits of this definition is that it can be used for all kinds of random variables (discrete, continuous, mixed, etc.). However, for this definition to be sound, there are two facts that need to be verified:

(a) Existence: It turns out that as long as X is integrable, a function ϕ with the above properties is guaranteed to exist. We already know that this is the case for discrete random variables: the conditional expectation as defined in the beginning of this section does have the desired properties. For general random variables, this is a nontrivial and deep result. It will be revisited later in this course.

(b) Uniqueness: It turns out that there is essentially only one function ϕ with the above properties. More precisely, any two functions with the above properties are equal with probability 1.

6.436J / 15.085J Fundamentals of Probability
Fall 2018