Math 525: Lecture 24

April 12, 2018

1 Conditional distribution and expectations

Let X and Y be random variables. If the event $\{X = x\}$ occurs with positive probability, we can define the *conditional distribution function* by

$$F_{Y|X}(y \mid x) = \mathbb{P}\left\{Y \leq y \mid X = x\right\}.$$

If Y is also integrable, we can also define the *conditional expectation* of Y given $X = x$:

$$\mathbb{E}\left[Y \mid X = x\right] = \frac{\mathbb{E}\left[Y I_{\{X=x\}}\right]}{\mathbb{P}\left\{X = x\right\}}.$$

We saw in a previous class that if in addition to being integrable, Y is discrete, the above simplifies:

$$\mathbb{E}\left[Y \mid X = x\right] = \frac{\sum_n y_n \mathbb{P}\left\{Y = y_n, X = x\right\}}{\mathbb{P}\left\{X = x\right\}} = \sum_n y_n \mathbb{P}\left\{Y = y_n \mid X = x\right\}.$$

However, when X has a continuous distribution, the event $\{X = x\}$ has probability zero. While it is possible to define $\mathbb{E}[Y \mid \cdot\,]$ in a general way that avoids this issue, doing so requires knowledge of some concepts from measure theory (e.g., Radon-Nikodym derivatives). Instead of working in the most general setting, we will instead tackle the case in which X and Y admit a joint density f_{XY}.

To motivate our definition of conditional distribution in this setting, suppose events of the form $\{x \leq X \leq x + h\}$ have positive probability (where $h > 0$) and consider

$$\mathbb{P}\left(Y \leq y \mid x \leq X \leq x + h\right).$$

Since X and Y admit a joint density,

$$\mathbb{P}\left(Y \leq y \mid x \leq X \leq x + h\right) = \frac{\mathbb{P}\left\{Y \leq y, x \leq X \leq x + h\right\}}{\mathbb{P}\left\{x \leq X \leq x + h\right\}} = \frac{\int_{-\infty}^{y} \int_{x}^{x+h} f_{XY}(u, v) du dv}{\int_{x}^{x+h} f_X(u) du}.$$

If h is small, we expect the above to be approximately

$$\frac{h \int_{-\infty}^{y} f_{XY}(x, v) dv}{h f_X(x)} = \frac{\int_{-\infty}^{y} f_{XY}(x, v) dv}{f_X(x)} = \int_{-\infty}^{y} \left(\frac{f_{XY}(x, v)}{f_X(x)} \right) dv.$$

This motivates the definition below.

Definition 1.1. Let X and Y be random variables which admit a joint density f_{XY}. We can define the *conditional density* of Y given $X = x$ by

$$f_{Y|X}(y \mid x) = \begin{cases} \dfrac{f_{XY}(x,v)}{f_X(x)} & \text{if } f_X(x) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The *conditional distribution function* of Y given $X = x$ is

$$F_{Y|X}(y \mid x) = \int_{-\infty}^{y} f_{Y|X}(v \mid x) dv.$$

The *conditional expectation* of Y given $X = x$ is

$$\mathbb{E}\left[Y \mid X = x\right] = \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x) dy.$$

Remark 1.2. If X and Y are independent in the above, the conditional density of Y given $X = x$ should be independent of X. Indeed, then

$$f_{Y|X}(y \mid x) = \frac{f_{XY}(x,y)}{f_X(x)} = f_Y(y).$$

Let's also verify that the conditional density works as desired. Indeed,

$$\begin{aligned} \int_a^b \overbrace{\left(\int_c^d f_{Y|X}(y \mid x) dy\right)}^{\mathbb{P}(c<Y\leq d|X=x)} f_X(x) dx &= \int_a^b \int_c^d \frac{f_{XY}(x,y)}{f_X(x)} f_X(x) dy dx \\ &= \int_a^b \int_c^d f_{XY}(x,y) dy dx \\ &= \mathbb{P}\left(a < X \leq b, c < Y \leq d\right). \end{aligned}$$

Definition 1.3. Suppose X is a discrete random variable or that X and Y admit a joint density. Suppose Y is integrable, so that the function ψ given by

$$\psi(x) = \mathbb{E}\left[Y \mid X = x\right]$$

is well-defined (if X is discrete, we can safely ignore points at which $\{X = x\}$). Then, the conditional expectation of Y given X is

$$\mathbb{E}\left[Y \mid X\right] \equiv \psi \circ X.$$

Unlike the previous definition, we have not specified the value of X in the conditional expectation. Therefore, unlike $\mathbb{E}\left[Y \mid X = x\right]$ (which is a scalar), $\mathbb{E}\left[Y \mid X\right]$ is itself a random variable. We now prove what is known as the *tower property of expectations*:[1]

[1]Actually, the tower property is more general, applying to the more general notion of conditional expectation we hinted at before.

Proposition 1.4. *Let X and Y be random variables which admit a joint density f_{XY}. Suppose $f_X \neq 0$. If Y is integrable, then $\mathbb{E}[Y \mid X]$ is also integrable and*

$$\mathbb{E}\left[\mathbb{E}\left[Y \mid X\right]\right] = \mathbb{E}Y.$$

Proof. First, suppose Y is positive. Then,

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}\left[Y \mid X\right]\right] &= \mathbb{E}\left[\psi(X)\right] \\
&= \int_{-\infty}^{\infty} \psi(x) f_X(x) dx \\
&= \int_{-\infty}^{\infty} \mathbb{E}\left[Y \mid X = x\right] f_X(x) dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x) dy f_X(x) dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y \frac{f_{XY}(x,y)}{f_X(x)} dy f_X(x) dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{XY}(x,y) dy dx \\
&= \mathbb{E}\left[Y\right].
\end{aligned}$$

If Y is not positive, we can split it into its positive and negative parts $Y = Y^+ - Y^-$. □

The tower property also holds for discrete random variables, in which case the intuition is very easy to understand:

Example 1.5. Consider two bags, labelled a and b. We pick bag a with probability P and bag b with probability $1-p$. We use the random variable X to indicate which bag we picked (i.e., X takes values in $\{a, b\}$). Both bags are filled with green and red balls. Define

$$p_a = \frac{\#\text{ of green balls in bag } a}{\text{total } \#\text{ of balls in bag } a} \qquad \text{and} \qquad p_b = \frac{\#\text{ of green balls in bag } b}{\text{total } \#\text{ of balls in bag } b}.$$

From the bag we have chosen, we pick a single ball. Let

$$Y = \begin{cases} 1 & \text{if we picked a green ball} \\ 0 & \text{if we picked a red ball.} \end{cases}$$

Then,

$$\mathbb{E}\left[Y \mid X = a\right] = p_a \qquad \text{and} \qquad \mathbb{E}\left[Y \mid X = b\right] = p_b.$$

Therefore,

$$\mathbb{E}\left[\mathbb{E}\left[Y \mid X\right]\right] = P\left(\mathbb{E}\left[Y \mid X = a\right]\right) + (1 - P)\,\mathbb{E}\left[Y \mid X = b\right] = P p_a + (1 - P)\, p_b = \mathbb{E}\left[Y\right].$$

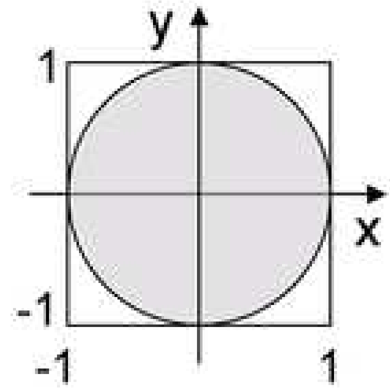

Figure 1: Circle with area π

2 Introduction to Monte Carlo methods

A Monte Carlo method is any algorithm that exploits the law of large numbers to compute a (determinstic) quantity. This is best motivated by example:

Example 2.1. Suppose we want to approximate the value of π. Recall that for a (Borel) function f,

$$\mathbb{E}\left[f(X)\right] = \frac{1}{b-a}\int_a^b f(x)dx \qquad \text{if } X \sim U[a,b].$$

The same is true in higher dimensions:

$$\mathbb{E}\left[f(X,Y)\right] = \frac{1}{d-c}\frac{1}{b-a}\int_c^d\int_a^b f(x,y)dxdy \qquad \text{if } X \sim U[a,b] \text{ and } Y \sim U[c,d].$$

Recall that π is the area of a unit circle. That is, letting $C = \{(x,y)\colon x^2+y^2 \leq 1\}$,

$$\pi = \int_{-1}^1\int_{-1}^1 I_C(x,y)dxdy.$$

In other words,

$$\mathbb{E}\left[I_C(X,Y)\right] = \frac{\pi}{4} \qquad \text{where} \qquad X,Y \sim U[-1,1].$$

Let $X_n, Y_n \sim U[-1,1]$. Then, by the strong law of large numbers,

$$\lim_{N\to\infty}\frac{I_C(X_1,Y_1)+\cdots+I_C(X_N,Y_N)}{N} = \frac{\pi}{4} \text{ a.s.}$$

This gives us a very simple algorithm to approximate π:

1. Pick a positive integer N and set S_N to zero.

2. Generate two (independent) random variables X and Y from the uniform distribution $U[-1,1]$. If $X^2+Y^2 \leq 1$, increment S_N by one.

3. Repeat step (2) n times. Then, $\pi/4 \approx S_N/N$.

Some MATLAB code for this procedure is given below:

```
% Generate (X, Y) pairs
X_Y_pairs = 2 * rand(2, N) - 1;

% Count which pairs are in the circle
S_N = sum(sum(pairs.^2) <= 1);

pi_approximate = 4 * (S_N / N);
```

With N equal to 10 million, a run of the code above yields $\pi_{\text{approximate}} = 3.142106$.

Even though we used 10 million samples in the above, our approximation is only accurate up to three significant digits! The slow convergence[2] is explained by the CLT, which says

$$\frac{1}{\sqrt{N}}\left(\frac{S_N}{N} - \frac{\pi}{4}\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma)$$

where $\sigma^2 = \text{Var}(I_C(X, Y))$. With a slight abuse of notation, we may write

$$\pi_{\text{approximate}} - \pi = 4\frac{S_N}{N} - \pi = O\left(\frac{1}{\sqrt{N}}\right)$$

to represent this fact.

[2]There are much more effective ways to compute π (e.g., Chudnovsky algorithm), but they do not have to do with Monte Carlo.