Lecture 9

Instructor: Quan Zhou

For more details about the materials covered in this note, see Chapters 10.2 and 10.3 of Resnick [6] and Chapter 4.1 of Durrett [3].

9.1 Conditional expectations

Definition 9.1. Consider a probability space $(\Omega, \mathcal{F}, \mathsf{P})$, a sub-$\sigma$-field $\mathcal{G} \subset \mathcal{F}$, and a random variable X such that $E|X| < \infty$. The conditional expectation of X given $\mathcal{G}$, denoted by $E[X \mid \mathcal{G}]$, is a random variable such that

(i) $E[X \mid \mathcal{G}]$ is $\mathcal{G}$-measurable;

(ii) for any $A \in \mathcal{G}$, we have $\int_A X d\,\mathsf{P} = \int_A E[X \mid \mathcal{G}] d\,\mathsf{P}$.

Any random variable that satisfies the above two properties is called a version of $E[X \mid \mathcal{G}]$. For two random variables X, Y defined on the same probability space, we often write $E[X \mid Y] = E[X \mid \sigma(Y)]$.

Theorem 9.1. *There exists a random variable that satisfies (i) and (ii) in Definition 9.1. Further, such a random variable is essentially unique, which means that any two versions of $E[X \mid \mathcal{G}]$ are equivalent almost surely.*

Proof. Here we only give the proof for a non-negative and integrable random variable $X \geq 0$.

(Existence.) It can be shown that (see also Theorem 5.6)

$$\nu(A) = \int_A X d\,\mathsf{P}, \qquad \forall\, A \in \mathcal{G}$$

defines a σ-finite measure on $(\Omega, \mathcal{G})$ (since $\nu(\Omega) < \infty$ by the integrability assumption). Let $\mathsf{P}_{|\mathcal{G}}$ be the restriction of P to $\mathcal{G}$; that is, $\mathsf{P}_{|\mathcal{G}}$ is a measure on $(\Omega, \mathcal{G})$ and $\mathsf{P}_{|\mathcal{G}}(A) = \mathsf{P}(A)$ for every $A \in \mathcal{G}$. Then, we have $\nu \ll \mathsf{P}_{|\mathcal{G}}$ and by Radon-Nikodym theorem, the derivative $d\nu / d\mathsf{P}_{|\mathcal{G}}$ is $\mathcal{G}$-measurable and

$$\int_A X d\,\mathsf{P} = \nu(A) = \int_A \frac{d\nu}{d\mathsf{P}_{|\mathcal{G}}} d\,\mathsf{P}_{|\mathcal{G}}, \qquad \forall\, A \in \mathcal{G}.$$

Because P agrees with $\mathsf{P}_{|\mathcal{G}}$ on $\mathcal{G}$, we have

$$\int_A \frac{d\nu}{d\mathsf{P}_{|\mathcal{G}}} d\,\mathsf{P}_{|\mathcal{G}} = \int_A \frac{d\nu}{d\mathsf{P}_{|\mathcal{G}}} d\,\mathsf{P}.$$

(This argument is not most rigorous. But one can prove it rigorously by starting from simple functions and then considering general non-negative functions.) So $d\nu/d\mathsf{P}_{|\mathcal{G}}$ is a version of $E[X \mid \mathcal{G}]$. Its existence follows from Radon-Nikodym theorem.

(Uniqueness.) By the uniqueness part of the Radon-Nikodym theorem, if there exists any other random variable, say Z, that satisfies properties (i) and (ii), it must be equal to $d\nu/d\mathsf{P}_{|\mathcal{G}}$, $\mathsf{P}_{|\mathcal{G}}$-a.e. But "$\mathsf{P}_{|\mathcal{G}}$-a.e." implies "$\mathsf{P}$-a.e.", which concludes the proof. □

Example 9.1. Consider a six-faced fair die. The sample space is given by $\Omega = \{1, 2, 3, 4, 5, 6\}$. Let P be the uniform probability measure on $(\Omega, \mathcal{F})$, where $\mathcal{F} = \mathcal{P}(\Omega)$, such that $\mathsf{P}(\{\omega\}) = 1/6$ for $\omega = 1, 2, \ldots, 6$. Let X be a random variable on $(\Omega, \mathcal{F})$ defined by $X(\omega) = \omega$. Consider a sub-σ-algebra $\mathcal{G}$ defined as

$$\mathcal{G} = \{\emptyset, \{1, 2, 3\}, \{4, 5, 6\}, \Omega\}.$$

Note that we can define a random variable $Y(\omega) = \mathbb{1}_{\{1,2,3\}}(\omega)$, which satisfies $\sigma(Y) = \mathcal{G}$. The conditional expectation $E[X \mid \mathcal{G}]$ is given by

$$E[X \mid \mathcal{G}](\omega) = \begin{cases} 2 & \text{if } \omega \in \{1, 2, 3\}, \\ 5 & \text{if } \omega \in \{4, 5, 6\}. \end{cases}$$

To verify this claim, we need to check the two conditions. The first one that $E[X \mid \mathcal{G}]$ is $\mathcal{G}$-measurable is obvious upon noticing that $E[X \mid \mathcal{G}]$ and Y should generate the same σ-algebra. To verify the second condition, we need to check the equality holds for all the four sets in $\mathcal{G}$. Here we only do it for the set $\{1, 2, 3\}$:

$$\int_{\{1,2,3\}} X d\mathsf{P} = \sum_{\omega=1}^{3} X(\omega)\mathsf{P}(\{\omega\}) = \frac{1}{6} \times (1 + 2 + 3) = 1,$$

$$\int_{\{1,2,3\}} E[X \mid \mathcal{G}] d\mathsf{P} = \sum_{\omega=1}^{3} E[X \mid \mathcal{G}](\omega)\mathsf{P}(\{\omega\}) = 2 \times (\frac{1}{6} + \frac{1}{6} + \frac{1}{6}) = 1.$$

Example 9.2. Consider a probability space $(\Omega, \mathcal{F}, \mathsf{P})$. Let $\Omega_1, \Omega_2, \dots$ be a countable partition of the entire sample space Ω ("partition" implies "disjoint") such that $\mathsf{P}(\Omega_i) > 0$ for each i. Define a sub-σ-algebra by

$$\mathcal{G} = \sigma(\Omega_1, \Omega_2, \dots).$$

Then, one can show that the conditional expectation of a random variable X given $\mathcal{G}$ is

$$E[X \mid \mathcal{G}](\omega) = \sum_{i \geq 1} \frac{\int_{\Omega_i} X d\mathsf{P}}{\mathsf{P}(\Omega_i)} \mathbb{1}_{\Omega_i}(\omega), \quad \text{a.s.}$$

Observe that equivalently this can be expressed as, almost surely,

$$E[X \mid \mathcal{G}](\omega) = \frac{E[X \mathbb{1}_{\Omega_i}]}{\mathsf{P}(\Omega_i)}, \quad \text{if } \omega \in \Omega_i.$$

This justifies why in elementary probability, we use the following formula to calculate the conditional expectation given any $A \in \mathcal{F}$,

$$E[X \mid A] = \frac{E[X \mathbb{1}_A]}{\mathsf{P}(A)}.$$

(In the above notation, $E[X \mid A]$ is a real number, not a random variable. We usually avoid using such notation in measure-theoretic probability.)

Remark 9.1. Let $Y \colon (\Omega, \mathcal{F}) \to (\Lambda, \mathcal{H})$. Consider a version of $E[X \mid \sigma(Y)]$, which by definition is a mapping from Ω to $\mathbb{R}$ and should be $\sigma(Y)$-measurable. By Proposition 3.5, there exists a function $h \colon (\Lambda, \mathcal{H}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $E[X \mid \sigma(Y)](\omega) = (h \circ Y)(\omega) = h(Y(\omega))$. This justifies why in statistics, we often use the notation $E[X \mid Y = y]$; it is defined as $E[X \mid Y = y] = h(y)$.

Remark 9.2. Consider $\mathbb{1}_{\{X \in A\}}$ for a random variable X and $A \in \mathcal{B}(\mathbb{R})$. Let Y be another random variable with an absolutely continuous distribution. From Remark 9.1, $\mathsf{P}(X \in A \mid Y = y) := E[\mathbb{1}_{\{X \in A\}} \mid Y = y] = h(y)$ for some measurable function h. Further, it can be shown that[1], for almost every y,

$$h(y) = \lim_{\delta \downarrow 0} \mathsf{P}(X \in A \mid Y \in (y - \delta, y + \delta]).$$

The right-hand side can be evaluated by using elementary formula for conditional probabilities. This yields a natural interpretation of $\mathsf{P}(X \in A \mid Y = y)$.

[1] This is quite non-trivial; see Probability and Measures by Billingsley.

Example 9.3. Let X, Y be independent standard normal random variables, and consider $\mathsf{P}(X \in A \mid X = Y)$. In light of Remark 9.2, we may want to interpret $\mathsf{P}(X \in A \mid X = Y)$ as the limit of $\mathsf{P}(X \in A \mid B_n)$ for some sequence of events $\{B_n\}_{n \geq 1}$ that converges to $\{X = Y\}$. This will be problematic, because the limit, even if it exists, largely depends on how we construct the sequence $\{B_n\}_{n \geq 1}$. For example, we can let $U = X - Y$ and $B_n^U = \{|U| < n^{-1}\}$; we can also let $V = X/Y$ and $B_n^V = \{|V - 1| < n^{-1}\}$. But $\lim_{n \to \infty} \mathsf{P}(X \in A \mid B_n^U)$ and $\lim_{n \to \infty} \mathsf{P}(X \in A \mid B_n^V)$ are unequal in general. (You can use the formula given in Proposition 9.4 to verify that the regular conditional distribution of $X \mid U = 0$ and $X \mid V = 1$ are actually different.) This is not too surprising upon observing that $\sigma(U) \neq \sigma(V)$. Whenever we do conditioning, we should think about the σ-algebra we are conditioning on. The two random variables $E[\mathbb{1}_{\{X \in A\}} \mid \sigma(U)]$ and $E[\mathbb{1}_{\{X \in A\}} \mid \sigma(V)]$ are very different. A similar example is given by the Borel-Kolmogorov paradox.

9.2 Properties of conditional expectations

For all results below, assume the probability space $(\Omega, \mathcal{F}, \mathsf{P})$ is given.

Proposition 9.1 (Basic properties of conditional expectation)**.** *Let X, Y be integrable random variables and $\mathcal{G} \subset \mathcal{F}$ be a given sub-σ-algebra.*

(i) *For $a, b \in \mathbb{R}$, $E[(aX + bY) \mid \mathcal{G}] = aE[X \mid \mathcal{G}] + bE[Y \mid \mathcal{G}]$, a.s.*

(ii) *If $X = c$ where $c \in \mathbb{R}$, then $E[X \mid \mathcal{G}] = c$, a.s.*

(iii) *If $X \geq Y$, then $E[X \mid \mathcal{G}] \geq E[Y \mid \mathcal{G}]$, a.s.*

(iv) *If $X \in \mathcal{G}$, then $E[X \mid \mathcal{G}] = X$, a.s.*

(v) *$E[X \mid \{\emptyset, \Omega\}] = E[X]$.*

(vi) *Law of total expectation: $E[\,E[X \mid \mathcal{G}]\,] = E[X]$.*

(vii) *Tower property: If $\mathcal{H}$ is another σ-algebra such that $\mathcal{H} \subset \mathcal{G} \subset \mathcal{F}$, then*

$$E[\,E[X \mid \mathcal{G}] \mid \mathcal{H}] = E[\,E[X \mid \mathcal{H}] \mid \mathcal{G}] = E[X \mid \mathcal{H}], \quad a.s.$$

(viii) *Suppose $E|XY| < \infty$ and $Y \in \mathcal{G}$. Then $E[XY \mid \mathcal{G}] = YE[X \mid \mathcal{G}]$, a.s.*

Remark 9.3. By part (vi), $E[\,E[X \mid Y]\,] = E[X]$ for any random variable Y, which is the non-measure theoretic version of the law of total expectation. Actually, part (vi) is just a special case of part (vii). Let $\mathcal{H} = \{\emptyset, \Omega\}$. Then, by part (v), $E[\,E[X \mid \mathcal{G}]] = E[\,E[X \mid \mathcal{G}] \mid \mathcal{H}] = E[X \mid \mathcal{H}] = E[X]$, a.s.

Proof of part (viii). We prove it using the definition of conditional expectation, i.e. we verify that $YE[X \mid \mathcal{G}]$ is a version of the conditional expectation of XY given $\mathcal{G}$ by checking the two conditions. The measurability part is easy. $YE[X \mid \mathcal{G}]$ is $\mathcal{G}$-measurable since both Y and $E[X \mid \mathcal{G}]$ are $\mathcal{G}$-measurable.

We also need to show $\int_A YE[X \mid \mathcal{G}]d\mathsf{P} = \int_A XY d\mathsf{P}$ for any $A \in \mathcal{G}$. We start by assuming $Y = \mathbb{1}_B$ for some $B \in \mathcal{G}$. Then,

$$\int_A YE[X \mid \mathcal{G}]d\mathsf{P} = \int_A \mathbb{1}_B E[X \mid \mathcal{G}]d\mathsf{P} = \int_{A\cap B} E[X \mid \mathcal{G}]d\mathsf{P}.$$

Since both A, B are in $\mathcal{G}$, we have $A \cap B \in \mathcal{G}$ and thus by the definition of conditional expectation,

$$\int_{A\cap B} E[X \mid \mathcal{G}]d\mathsf{P} = \int_{A\cap B} X d\mathsf{P} = \int_A \mathbb{1}_B X d\mathsf{P} = \int_A XY d\mathsf{P}.$$

It is straightforward to repeat the above calculations for simple functions. Next, assume that both X, Y are non-negative and apply MCT (details omitted here). Finally, by writing $X = X^+ - X^-$ and $Y = Y^+ - Y^-$ (details omitted again), we can prove the proposition for any integrable X, Y such that XY is also integrable and $Y \in \mathcal{G}$. □

Proof of the remaining part(s). Try it yourself. □

Proposition 9.2 (Conditional expectation and independence)**.** *Let X, Y, Z be integrable random variables and $\mathcal{G} \subset \mathcal{F}$ be a given sub-σ-algebra.*

(i) If $\sigma(X)$ and $\mathcal{G}$ are independent, then $E[X \mid \mathcal{G}] = E[X]$, a.s.

(ii) Suppose X, Y are independent, and ϕ is a Borel function such that $E|\phi(X, Y)| < \infty$. Define a function f by letting $f(x) = E[\phi(x, Y)]$ for each $x \in \mathbb{R}$. Then, $E[\phi(X, Y) \mid X] = f(X)$, a.s.

(iii) If $\sigma(X, Y)$ is independent of $\sigma(Z)$, $E[Y \mid X, Z] = E[Y \mid X]$, a.s.

Proof of part (i). Try it yourself. □

Proof of part (ii). See Example 4.1.7 in Durrett [3] and part (12) in §10.3 of Resnick [6]. □

Sketch of proof of part (iii). Let $W = E[Y \mid X]$. We show that W is a version of $E[Y \mid X, Z]$ by verifying the two conditions. The measurability part is easy. The second condition is that $E[W \mathbb{1}_A] = E[Y \mathbb{1}_A]$ for every $A \in \sigma(X, Z)$. We begin by considering measurable rectangle set $B_1 \times B_2 \in \mathcal{B}(\mathbb{R}^2)$. Using the independence assumption, one can show that

$$E[W \mathbb{1}_{B_1 \times B_2}(X, Z)] = E[Y \mathbb{1}_{B_1 \times B_2}(X, Z)].$$

Define $\mathcal{L} = \{B \in \mathcal{B}(\mathbb{R}^2) \colon E[W \mathbb{1}_B(X, Z)] = E[Y \mathbb{1}_B(X, Z)]\}$. Show that $\mathcal{L}$ is a λ-system and use Dynkin's theorem to conclude the proof. □

Proposition 9.3 (Limits of conditional expectation)**.** *Let X and $\{X_n\}$ be integrable random variables and $\mathcal{G} \subset \mathcal{F}$ be a given sub-σ-algebra.*

(i) MCT: If $0 \leq X_n \uparrow X$, then $E[X_n \mid \mathcal{G}] \uparrow E[X \mid \mathcal{G}]$, a.s.

(ii) DCT: If $X_n \to X$ and $|X_n| \leq Z$ for some integrable random variable Z, then $E[X \mid \mathcal{G}] = \lim_{n \to \infty} E[X_n \mid \mathcal{G}]$, a.s.

Proof. See the textbook. □

9.3 Conditional probability

Definition 9.2. Consider the probability space $(\Omega, \mathcal{F}, \mathsf{P})$, a sub-$\sigma$-field $\mathcal{G} \subset \mathcal{F}$, and a random variable $X \colon (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $E|X| < \infty$. The conditional probability $\mathsf{P}(X \in A \mid \mathcal{G})$ for any $A \in \mathcal{B}(\mathbb{R})$ is defined as

$$\mathsf{P}(X \in A \mid \mathcal{G}) = E[\mathbb{1}_{\{X \in A\}} \mid \mathcal{G}].$$

Remark 9.4. By the above definition and Proposition 9.1 (iii), we have $\mathsf{P}(X \in A \mid \mathcal{G}) \in [0, 1]$ a.s. Further, $\mathsf{P}(X \in \emptyset \mid \mathcal{G}) = 0$ and $\mathsf{P}(X \in \Omega \mid \mathcal{G}) = 1$, a.s. Let $A_1, A_2, \dots$ be a sequence of disjoint Borel sets. Then,

$$\begin{aligned}\sum_{n=1}^{\infty} \mathsf{P}(X \in A_n \mid \mathcal{G}) &= \sum_{n=1}^{\infty} E[\mathbb{1}_{\{X \in A_n\}} \mid \mathcal{G}] \\ &= \lim_{n \to \infty} \sum_{i=1}^{n} E[\mathbb{1}_{\{X \in A_i\}} \mid \mathcal{G}] \\ &= \lim_{n \to \infty} E[\mathbb{1}(X \in \cup_{i=1}^{n} A_i) \mid \mathcal{G}] \\ &= E[\mathbb{1}(X \in \cup_{i=1}^{\infty} A_i) \mid \mathcal{G}] \\ &= \mathsf{P}(X \in \cup_{n=1}^{\infty} A_n \mid \mathcal{G}),\end{aligned}$$

almost surely. Note that the second last line follows from the MCT for conditional expectation. It is tempting to jump to the conclusion that a.s. $\mathsf{P}(X \in \cdot \mid \mathcal{G})$ is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. But we cannot. For any given sequence $\{A_n\}$ of disjoint sets, the countable additivity may fail to hold on a P-null set (i.e. a set with probability 0). Since there could be uncountably many such sequences, the union of all these P-null sets may have positive probability.

Theorem 9.2. *Consider the setting of Definition 9.2. There always exists a function $p : \Omega \times \mathcal{B}(\mathbb{R}) \to [0, 1]$, which is called a regular conditional distribution of X given $\mathcal{G}$, such that*

(i) for each $A \in \mathcal{B}(\mathbb{R})$, the function $p(\cdot, A)$ is a version of $\mathsf{P}(X \in A \mid \mathcal{G})$;

(ii) for P-almost every $\omega \in \Omega$, the function $p(\omega, \cdot)$ is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Proof. See Durrett [3, §4.1.3]. □

Remark 9.5. A measurable function from $(\Omega, \mathcal{F}, \mathsf{P})$ to $(\Lambda, \mathcal{G})$ is called a random element. The above theorem is not true if X is a random element, and there are explicit counterexamples where the regular conditional distribution does not exist.

Proposition 9.4. *Let $Z = (X, Y)\colon (\Omega, \mathcal{F}, \mathsf{P}) \to (\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ be a random vector with density $f_Z = d(\mathsf{P} \circ Z^{-1})/dm^2$. Define $f_Y(y) = \int_{\mathbb{R}} f_Z(x, y) m(dx)$ and $f_{X|Y}(x, y) = f_Z(x, y)/f_Y(y)$. Then,*

$$p(\omega, A) = \int_A f_{X|Y}(x, Y(\omega)) m(dx), \qquad \forall \omega \in \Omega,\ A \in \mathcal{B}(\mathbb{R}).$$

is the regular conditional distribution of X given $\sigma(Y)$. In other words, the regular conditional distribution of X given $Y = y$ has density $f_{X|Y}(\cdot, y)$.

Proof. We check the two conditions. First, fix an arbitrary $A \in \mathcal{B}(\mathbb{R})$ and consider the mapping $\omega \mapsto p(\omega, A)$. By Fubini's theorem[2] and composition theorem (Proposition 3.4), this mapping is $\sigma(Y)$-measurable. To show that $\omega \mapsto p(\omega, A)$ is a version of $\mathsf{P}(X \in A \mid Y)$, it suffices to prove that for any $B \in \mathcal{B}(\mathbb{R})$, we have

$$\int_{Y^{-1}(B)} p(\omega, A)\mathsf{P}(d\omega) = \int_{Y^{-1}(B)} \mathbb{1}_{\{X \in A\}}\mathsf{P}(d\omega) = \mathsf{P}(X \in A, Y \in B).$$

Observe that f_Y is the marginal density function of the random variable Y; that is, $f_Y = d(\mathsf{P} \circ Y^{-1})/dm$. So, the change-of-variable formula,

$$\begin{aligned}
&\int_{Y^{-1}(B)} \left\{ \int_A f_{X|Y}(x, Y(\omega))m(dx) \right\} \mathsf{P}(d\omega) \\
&= \int_B \left\{ \int_A \frac{f_Z(x,y)}{f_Y(y)} m(dx) \right\} (\mathsf{P} \circ Y^{-1})(dy) \\
&= \int_B \frac{1}{f_Y(y)} \left\{ \int_A f_Z(x,y)m(dx) \right\} (\mathsf{P} \circ Y^{-1})(dy) \\
&= \int_B \left\{ \int_A f_Z(x,y)m(dx) \right\} m(dy) \\
&= \mathsf{P}(Z \in A \times B) = \mathsf{P}(X \in A, Y \in B).
\end{aligned}$$

Second, the mapping $A \mapsto p(\omega, A)$ is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ by Theorem 5.6. This concludes the proof. □

References

[1] Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2017.

[2] Dennis D. Cox. *The Theory of Statistics and Its Applications*. Unpublished.

[2] In Theorem 6.2, the mapping $\omega_1 \mapsto \int_{\Omega_2} f(\omega_1, \omega_2)\mu_2(d\omega_2)$ has to be measurable since otherwise Fubini's theorem may not make sense. This measurability result is indeed part of the Fubini's theorem; see Resnick [6] for details.

[3] Rick Durrett. *Probability: Theory and Examples*, volume 49. Cambridge university press, 2019.

[4] Achim Klenke. *Probability theory: a comprehensive course.* Springer Science & Business Media, 2013.

[5] Michael A Proschan and Brett Presnell. Expect the unexpected from conditional expectation. *The American Statistician*, 52(3):248–252, 1998.

[6] Sidney Resnick. *A Probability Path.* Springer, 2019.