4 Conditional Expectations

Conditional Probability: Discrete Case

- Let $X : (\Omega, \mathcal{F}, P) \to (D_X, 2^{D_X}, P_X)$ and $Y : (\Omega, \mathcal{F}, P) \to (D_Y, 2^{D_Y}, P_Y)$ be two random variables with joint PMF $p_{XY}(x, y)$.
- When X takes value $x_i \in D_X$, the conditional probability that Y takes value $y_j \in D_Y$ is

$$P(Y = y_j \mid X = x_i) = \frac{P(X = x_i, Y = y_j)}{P(X = x_i)} = \boxed{\frac{p_{XY}(x_i, y_j)}{p_X(x_i)} =: p_{Y|X}(y_j \mid x_i)}$$

- Using this notation, for any $B \in 2^{D_Y}$ and $A \in \mathcal{F}$, we have

$$P(Y \in B \mid A) = \frac{P(\{\omega : Y(\omega) = B, \omega \in A\})}{P(A)} = \frac{P(Y^{-1}(B) \cap A\})}{P(A)}$$

If we take A as $X = x_i$, then

$$P(Y \in B \mid X = x_i) = \frac{P(Y \in B, X = x_i)}{P(X = x_i)} = \frac{\sum_{y_j \in B} P(X = x_i, Y = y_j)}{P(X = x_i)} = \sum_{y_j \in B} p_{Y|X}(y_j \mid x_i)$$

If we take $A \in 2^{D_X}$, then

$$P(Y \in B \mid X \in A) = \frac{P(Y \in B, X \in A)}{P(X \in A)} = \frac{\sum_{x_i \in A} \sum_{y_j \in B} p_{XY}(x_i, y_j)}{\sum_{x_i \in A} p_X(x_i)}$$

Note that the above is NOT equal to $\sum_{x_i \in A} \sum_{y_j \in B} p_{Y|X}(y_j \mid x_i)$. Try to find a counter-example by yourself.

Conditional Probability: Continuous Case

- For continuous random variable $X : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$, let us consider

$$(\star) = P(Y \in B \mid x < X \leq x + \Delta x) = \frac{P(Y \in B, X \in (x, x + \Delta x])}{P(X \in (x, x + \Delta x])}$$

Then take the limit as $\Delta x \to 0$, we get

$$\lim_{\Delta x \to 0} (\star) = \frac{\int_B f_{XY}(x, y) \Delta x dy}{f_X(x) \Delta x} = \frac{\int_B f_{XY}(x, y) dy}{f_X(x)} = \int_B \frac{f_{XY}(x, y)}{f_X(x)} dy$$

Therefore, we define

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

as the **conditional PDF** of Y given X.

Conditional Expectation: Discrete Case

- Then given any $x_i \in D_X$, the distribution as well as the expectation of Y changes from

$$E(Y) = \sum_{y_j \in D_Y} y_j p_Y(y_j) \quad \text{to} \quad E(Y \mid X = x_i) = \sum_{y_j \in D_Y} y_j p_{Y|X}(y_j \mid x_i)$$

  Therefore, $E(Y \mid X = x_i)$ is the conditional expectation of Y given $X = x_i$, which is a **real number**.

- Therefore, it makes sense to consider a function $\psi : \mathbb{R} \to \mathbb{R}$ defined by:

$$\forall x \in D_X : \psi(x_i) = E(Y \mid X = x_i).$$

  Hence, $\psi(X) = E(Y \mid X)$ is called the **conditional expectation** of Y given X, which is a new **random variable** $E(Y \mid X)(\omega) = E(Y \mid X = X(\omega))$.

- In fact, this new random variable $E(Y \mid X)$ has the following properties

  (1) it is measurable w.r.t. the σ-field $\sigma(X) = \{X^{-1}(A) : A \in 2^{D_X}\}$

  (2) for every $x_i \in D_X$, we have

$$\int_{\{X=x_i\}} E(Y \mid X) dP = \int_{\{X=x_i\}} Y dP$$

  The first property holds because for any $\omega \in X^{-1}(\{x_i\})$, $E(Y \mid X)(\omega)$ has the same value. The second property comes from the following calculation

$$\begin{aligned}\int_{\{X=x_i\}} E(Y \mid X) dP =& E(Y \mid X = x_i) P(X = x_i) = \sum_{y_j \in D_Y} y_j P(Y = y_j \mid X = x_i) P(X = x_i) \\ =& \sum_{y_j \in D_Y} y_j P(Y = y_j, X = x_i) = \int_{\{X=x_i\}} Y dP\end{aligned}$$

  By denoting $Z = E(Y \mid X)$, this can be further simplied as

$$E(Z \mathbf{1}_{\{X=x_1\}}) = E(Y \mathbf{1}_{\{X=x_1\}})$$

- For example, $E(\mathbf{1}_A \mid \mathbf{1}_B)(\omega) = \begin{cases} P(A \mid B) & \text{if } \omega \in B \\ P(A \mid B^c) & \text{if } \omega \notin B \end{cases}$

- The essence of conditioning or conditional expectation can be interpreted from the information point of view. Suppose that X and Y are both defined on $(\Omega, \mathcal{F}, P)$ and are not independent. Essentially, they are two perspectives of the probability space! Therefore, observing something concrete event will of course give you more information. However, just having the new perspective Y the thing itself will also give you new information! This information is actually $\sigma(Y)$. This leads to the basic idea of our later development that **information is actually a sub σ-field $\mathcal{G} \subseteq \mathcal{F}$.**

- Suppose that $\Omega = \{1, 2, \ldots, 6\}, \mathcal{F} = 2^\Omega$ and $P(k) = 1/6, \forall k$, which is the model for die toss. Then the information "*whether an even number was thrown*" is encoded as the σ-field $\mathcal{G} = \{\emptyset, \{2, 4, 6\}, \{1, 3, 5\}, \Omega\}$.

Properties of Conditional Expectation

- Let $X : (\Omega, \mathcal{F}, P) \to (D_X, 2^{D_X}, P_X)$ and $Y : (\Omega, \mathcal{F}, P) \to (D_X, 2^{D_X}, P_Y)$. Then

Theorem: Law of Total Expectation

$$\boxed{E(Y) = E(E(Y|X))}$$

This theorem is also called the law of total expectation, the law of iterated expectation (LIE), the tower rule, and the smoothing theorem.

Proof: We only prove the discrete case, but it holds for any random variable.

$$\begin{aligned} E(E(Y|X)) &= \sum_{x_i \in D_X} E(Y|X = x_i) p_X(x_i) = \sum_{x_i \in D_X} \left(\sum_{y_i \in D_Y} y_j p_{Y|X}(y_j \mid x_i) \right) p_X(x_i) \\ &= \sum_{x_i \in D_X} \sum_{y_i \in D_Y} y_j p_{XY}(x_i, y_j) = \sum_{y_i \in D_Y} y_j p_Y(y_j) \\ &= E(Y) \end{aligned}$$

- Let $A_1 \dot\cup A_2 \dot\cup \cdots$ be a countable partition of Ω. Then $\boxed{E(X) = \sum_{i=1}^{\infty} E(X \mid A_i) P(A_i)}$.

- Let $X : (\Omega, \mathcal{F}, P) \to (D_X, 2^{D_X}, P_X)$ and $Y = h(X)$. Then $\boxed{E(Y \mid X) = h(X)}$.
Proof: It suffices to show that $E(h(X) \mid X = x_i) = h(x_i)$ for any $x_i \in D_X$. First,

$$E(h(X) \mid X = x_i) = \sum_{y_j \in D_Y} y_j p_{Y|X}(y_j \mid x_j) = \sum_{y_j \in D_Y} y_j P(Y = y_j \mid X = x_j)$$

Furthermore,

$$P(Y = y_j \mid X = x_j) = \begin{cases} 1 & \text{if } y_j = h(x_i) \\ 0 & \text{otherwise} \end{cases}$$

Therefore, we have $E(h(X) \mid X = x_i) = h(x_i)$.

- Let $X : (\Omega, \mathcal{F}, P) \to (D_X, 2^{D_X}, P_X)$ and $Y : (\Omega, \mathcal{F}, P) \to (D_X, 2^{D_X}, P_Y)$. Then

$$\boxed{E(g(X)Y \mid X) = g(X)E(Y \mid X)}$$

Proof: It suffices to show that $E(g(x_i)Y \mid X = x_i) = g(x_i)E(Y \mid X = x_i)$ for any $x_i \in D_X$, which is obvious as $g(x_i)$ is just a constant.

Examples of Conditional Expectations

▶ **Example of the Computation of $\mathbf{E(X \mid Y)}$**

A clinic has N people come to take COVID-19 vaccines, where N has the Poisson distribution with parameter λ. Each person has probability p to be immune. Let K be the total number of people get immune from the clinic. Find $E(K \mid N)$.

We are given that

$$f_N(n) = \frac{\lambda^n}{n!}e^{-\lambda}, \quad f_{K|N}(k \mid n) = \binom{n}{k}p^k(1-p)^{n-k}$$

Therefore,

$$\psi(n) = E(K \mid N = n) = \sum_k k f_{K|N}(k \mid n) = pn$$

Thus, we have

$$E(K \mid N) = \psi(N) = pN \text{ and } E(K) = E(E(K \mid N)) = pE(N) = p\lambda$$

Question: what is $E(N \mid K)$?

▶ **Connection to Mean Square Error Estimation**

Suppose that we are given X and Y with joint PDF f_{XY}. We observe Y and want to determine an estimate $\hat{X}$ of X, where $\hat{X} = h(Y)$, so as to minimize $E((X - \hat{X})^2)$.

The solution is as follows.

$$\begin{aligned}
&E((X-\hat{X})^2)\\
=&E(((X - E(X \mid Y)) + (E(X \mid Y) - \hat{X}))^2)\\
=&E((X - E(X \mid Y))^2) + E((\underbrace{E(X \mid Y) - \hat{X}}_{=:g(Y)})^2) + 2E((X - E(X \mid Y))(\underbrace{E(X \mid Y) - \hat{X}}_{=:g(Y)}))\\
=&E((X - E(X \mid Y))^2) + E(g^2(Y)) + 2E((X - E(X \mid Y))g(Y))
\end{aligned}$$

For the last term of the above, have

$$E((X - E(X \mid Y))g(Y)) = E(Xg(Y)) - E(E(X \mid Y)g(Y)) = 0$$

Therefore, $E((X - \hat{X})^2)$ is minimized when $h(Y) = E(X \mid Y)$.

The above also has a nice geometric interpretation. Essentially, it says that $(X - E(X \mid Y))$ and $g(Y)$ are *orthogonal* for any $g(\cdot)$. Therefore, the best estimate has to be $X - (X - E(X \mid Y)) = E(X \mid Y)$.

General Definition of Conditioning

- When an arbitrary event $B \in \mathcal{F}$ in the probability space is given, the expectation of the random variable will be changed. Essentially, B induces a new probability measure

$$P(? \mid B) = \frac{P(?, B)}{P(B)} = \int_B \mathbf{1}_? \frac{dP}{P(B)}$$

Definition: Conditioning on an Event

For any integrable random variable $X : \Omega \to \mathbb{R}$ and any event $B \in \mathcal{F}$ such that $P(B) \neq 0$, the **conditional expectation** of X given B is defined by

$$E(X \mid B) = \frac{1}{P(B)} \int_B X dP$$

For example, for the indicator random variable $\mathbf{1}_A : \Omega \to \{0, 1\}$, we have

$$E(\mathbf{1}_A \mid B) = \frac{1}{P(B)} \int_B \mathbf{1}_A dP = \frac{1}{P(B)} \int_{A \cap B} dP = \frac{1}{P(B)} P(A \cap B) = P(A \mid B)$$

- Given a random variable $X : \Omega \to \mathbb{R}$, recall that $\sigma(X) = \{X^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\}$ is the its generated σ-field. For the discrete case, for each $x_i \in D_X$, $\{X = x_i\}$ is an event in $\sigma(X)$, and $E(Y \mid X = x_i)$ is the conditional expectation given $\{X = x_i\}$. Therefore, we can think $E(Y \mid X)$ as a random variable whose generated σ-field is also $\sigma(X)$. Furthermore, $\int_{\{X=x_i\}} E(Y \mid X) dP = \int_{\{X=x_i\}} Y dP$ suggests that, for any $B \in 2^{D_X}$, we also have $\int_B E(Y \mid X) dP = \int_B Y dP$. We can generalize this requirement as follows.

Definition: Conditioning on an Random Variable

Let $Y : \Omega \to \mathbb{R}$ be an integrable random variable and $X : \Omega \to \mathbb{R}$ be an arbitrary random variable. Then the **conditional expectation** of Y given X is as a new random variable $E(Y \mid X)$ such that it is $\sigma(X)$-measurable, and for any $B \in \sigma(X)$

$$\int_B E(Y \mid X) dP = \int_B Y dP$$

- In fact, we can replace $\sigma(X)$ above by any σ-field $\mathcal{G} \subseteq \mathcal{F}$.

Definition: Conditioning on an σ-Field

Let $X : \Omega \to \mathbb{R}$ be an integrable random variable and $\mathcal{G}$ be a σ-field contained in $\mathcal{F}$. Then the **conditional expectation** of X given $\mathcal{G}$ is as a new random variable $E(Y \mid \mathcal{G})$ such that it is $\mathcal{G}$-measurable, and for any $B \in \mathcal{G}$

$$\int_B E(Y \mid \mathcal{G}) dP = \int_B Y dP$$

- You may ask why $E(Y \mid \mathcal{G})$ exists or it is unique. The answer is YES: $E(Y \mid \mathcal{G})$ always exists and unique due to the Radon-Nikodym Theorem, which will not be discussed here.

More About the Conditional Expectation

Let $(\Omega, \mathcal{F}, P)$ be a probability space, $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-field and X, Y be random variables with $E(|X|), E(|Y|) < \infty$. Then the conditional expectations have the following properties:

- **Linearity:** $E(aX + bY \mid \mathcal{G}) = aE(X \mid \mathcal{G}) + bE(Y \mid \mathcal{G})$.
- **Independence:** If X is independent of $\mathcal{G}$, then $E(X \mid \mathcal{G}) = E(X)$.
- **Tower property:** If $\mathcal{H} \subseteq \mathcal{G}$, then $E(E(X \mid \mathcal{G}) \mid \mathcal{H}) = E(X \mid \mathcal{H})$.
- **Taking out what is known-1:** If X is $\mathcal{G}$-measurable, then $E(X \mid \mathcal{G}) = X$.
- **Taking out what is known-2:** If X is $\mathcal{G}$-measurable, then $E(XY \mid \mathcal{G}) = XE(Y \mid \mathcal{G})$.
- For $\mathcal{G} = \{\emptyset, \Omega\}$, we have $E(X \mid \mathcal{G}) = E(X)$.
  Proof: We take verify that, by taking $E(X \mid \mathcal{G})$ to be constant random variable $E(X)$, it satisfies the requirements of conditional expectation. Specifically, we have $\int_\Omega E(X \mid \mathcal{G})dP = \int_\Omega E(X)dP = E(X) = \int_\Omega XdP$. Furthermore, such $E(X \mid \mathcal{G})$ is uniqiue.
- $E(E(X \mid \mathcal{G})) = E(X)$.
  Proof: By taking $\Omega \in \mathcal{G}$, we have $E(E(X \mid \mathcal{G})) = \int_\Omega E(X \mid \mathcal{G})dP = \int_\Omega XdP = E(X)$.
- If $B \in \mathcal{G}$, then $E(E(X \mid \mathcal{G}) \mid B) = E(X \mid B)$.
  Proof: $E(E(X \mid \mathcal{G}) \mid B) = \frac{\int_B E(X|\mathcal{G})dP}{P(B)} = \frac{\int_B XdP}{P(B)} = E(X \mid B)$.