Conditional Probability vs Conditional Expectation

Let \( A \subseteq \mathcal{A} \) and \( P(A) \neq 0 \). The conditional probability of \( B \subseteq \mathcal{A} \) given \( A \) is defined to be \( \frac{P(A \cap B)}{P(A)} \) and is denoted by \( P(B | A) \). (For \( P(A) = 0 \), \( P(\cdot | A) \) is undefined.) For fixed \( A \), the set function \( P(\cdot | A) \) defined on \( \mathcal{A} \) is a probability measure.

So, from definition of conditional probabilities, we obtain the multiplicative formula for probabilities:

\[
P(A \cap B) = P(A) P(B | A)
\]

More generally, if \( A_1, A_2, \ldots, A_n \) are events such that \( P\left(\bigcap_{i=2}^{n} A_i\right) \neq 0 \), then

\[
P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1) P(A_2 | A_1) \ldots P(A_n | \bigcap_{i=2}^{n-1} A_i).
\]

The following law of total probability is particularly helpful for computing probabilities of complicated events.
Let \( \{A_1, A_2, \ldots, A_n\} \) be a measurable partition of \( \Omega \), \( A_i \in \mathcal{A}, \ i = 1, 2, \ldots, n \), the \( A_i \)'s are disjoint and \( \bigcup_{i=1}^{n} A_i = \Omega \). Assuming that \( P(A_i) > 0 \), \( i = 1, 2, \ldots, n \), then for any \( B \subseteq \Omega \),

\[
P(B) = \sum_{i=1}^{n} P(B \mid A_i) P(A_i)
\]

As a consequence, we get Bayes' formula: if \( P(B) > 0 \)

then

\[
P(A_i \mid B) = \frac{P(A_i) P(B \mid A_i)}{\sum_{j=1}^{n} P(A_j) P(B \mid A_j)} \quad \text{for any } i = 1, 2, \ldots, n
\]

Independence of events

For \( A, B \in \mathcal{A} \) with \( P(A) \neq 0 \) and \( P(B) \neq 0 \), it is intuitive that "\( A \) is independent of \( B \)" (with respect to \( \mathcal{A} \) and \( P \)) when

\[
P(A \mid B) = P(A)
\]

and similarly "\( B \) is independent of \( A \)" when

\[
P(B \mid A) = P(B)
\]

In both cases,

\[
P(A \cap B) = P(A)P(B)
\]

The arbitrary collection \( \{A_i : i \in I\} \) of \( \mathcal{A} \) sets is said to be independent if for any finite \( J \subseteq I \),

\[
P\left( \bigcap_{i \in J} A_i \right) = \prod_{i \in J} P(A_i)
\]
The independence of $A_1, A_2, \ldots, A_n$ implies that any two events $A_i$ and $A_j$ are independent (pairwise independent) but the converse does not hold.

Finally, for $A, B, C \in \mathcal{A}$, we say that $A$ and $B$ are independent given $C$ if

$$P(A \cap B | C) = P(A | C) P(B | C)$$

The general concept of conditional independence appears naturally in the context of Markov processes.

Next we will discuss the concept of conditional distributions.

Recall $P_A(\cdot): \mathcal{A} \rightarrow [0, 1]$, $P_A(B) = P(B | A)$

is a probability measure on $\mathcal{A}$ and is called the conditional probability measure given $A$.

In applications, when several random variables are involved, often we are interested in computing

$$P(Y \in A | X = x)$$

for event $A$ in the range of $Y$.

This set function can be associated with a distribution function $F(y | x) = P[Y \leq y | X = x]$. It is the conditional distribution of $Y$ given $X = x$ in this context.
function is well defined for $P[\sum x = 0] \neq 0$.

For example, suppose that $X$ is discrete with support $\{x_1, x_2, \ldots, x_n\}$ and $P[X = x_n] > 0, n \geq 1$ and $\sum_{n=1}^{\infty} P[X = x_n] = 1$, then $F(x|x_n)$ is the distribution of $Y$ after observing the value $x_n$.

Before observing $X$, $P[Y \in A | X]$ is a random variable defined as

$$P(Y \in A | X) = \sum_{n=1}^{\infty} P(A | B_n) B_n(x),$$

where $B_n = \{x: X = x_n\}$. Note $\{B_n, n \geq 1\}$ form a partition of $\Omega$.

When $X$ is continuous r.v. (so that $P(X = x) = 0 \forall x$), the definition of conditional distribution is a little trickier. (Example of coin tossing)

For computational purposes, when the pair of r.v.s $(X, Y)$ has a joint density $f(x, y)$ then

$$f(y|x) = \int f(z|x)dz,$$

where the conditional density function is

$$f(y|x) = \frac{f(x, y)}{f_x(x)} \text{ for } f_x(x) \neq 0.$$

it is
The random variables $X_1, \ldots, X_n$ are said to be (mutually) independent if
\[ P(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} P(X_i \in A_i), \forall A_i \in \mathcal{B}. \]

The independence of an arbitrary collection of random variables is defined as
\[ f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_i(x_i), \forall (x_1, \ldots, x_n) \in \mathbb{R}^n, \]
where $f$ is the joint mass (density) probability function of the $X_1, \ldots, X_n$ and $f_i$ is the marginal mass (density) probability function of $X_i$.

Sums of independent random variables appear often in studies of stochastic processes.

Let $X$ and $Y$ be two independent discrete random variables with values in $\{0, 1, 2, \ldots\}$. The distribution of $Z = X + Y$ is completely determined by the mass probability function $f_Z$.
\[ f_Z(n) = \varnothing(\{2n\}) = P(X + Y = n), \quad n \geq 0. \]
Now, for fixed n,
\[ (x + y = n) = \bigcup_{k=1}^{n} (x = k, \ y = n-k) . \]

Since the events \( E_k : X(A) = k, \ Y(A) = n-k, 1 \leq k \leq n \) are disjoint, we have
\[
P(x + y = n) = \sum_{k=0}^{n} P(x = k, \ y = n-k) = \sum_{k=0}^{n} P(x = k) P(y = n-k) .
\]

By independence,
\[
f_z(z) = \int_{-\infty}^{\infty} f_x(x) f_y(z-x) dx, \ x \in \mathbb{R}.
\]

This operation is called convolution.

Let consider few aspects of conditional independence.

Consider two random variables \( X \) and \( Y \), defined on \((\Omega, \mathcal{A}, P)\). Suppose that \( X \) is discrete on \( \mathbb{N} \) with range \{\( x_n, n \geq 1 \}\). The variable \( X \) induces a measurable partition (finite or countable)
\[
\mathcal{B} = \mathcal{A} ,
\]
\[
B_n = \{ \omega : X(\omega) = x_n \}, n \geq 1.
\]

When \( X = x_n \), we might be interested in \( P(A|X=x_n) \) for some set \( A \neq \emptyset \) of \( x_n \).
Before observing $X$, the conditional prob. of the event: A given $X$ is a random variable defined as

$$P(A|X)(x) = \sum_{n \geq 1} P(A \cap D_n) I_{D_n}(x)$$

If $Y$ is a r.v. with finite range $\{y_1, y_2, \ldots, y_m\}$, then

$$E(Y) = \sum_{i} y_i \cdot P(B_i) \quad B_i = \{Y = y_i\}$$

$$E(Y|X = x) = E(Y|D_n) = \sum_{i} y_i \cdot P(B_i|D_n)$$

$$= \frac{1}{P(D_n)} \sum_{i} y_i \cdot P(B_i \cap D_n)$$

$$= \frac{E(Y|D_n)}{P(D_n)}$$

In general, if the extended r.v. $Y$ whose expection exists, then $E(Y|D)$ exists for $D \in A$ with $P(D) > 0$, where

$$E(Y|D) = \int_{D} y \cdot dP_D(y)$$

and $P_D(\cdot)$ is the conditional probability measure on $A$ defined by

$$P_D(A) = P(A|D), \quad A \in \mathcal{A}$$
the Radon-Nikodym theorem guarantees the existence of the conditional expectation function and also says that the function is unique. Thus if there exists another function, say \( x(d) \) that also satisfies the above definition of conditional expectation, then we have \( E[Y|D] \equiv x(d) \wrt P \).

Some useful properties of conditional expectation:

1. \( E(1|D) \) is increasing and linear:

\[ X \leq Y \Rightarrow E(X|D) \leq E(Y|D) \quad (a.s) \]

where a.s.

for \( x, \beta < R \),

\[ E(\alpha X + \beta Y|D) = \alpha E(X|D) + \beta E(Y|D) \quad (a.s) \]

(5) For \( D = \{ \emptyset, \Omega \} \), \( E(X|D) = E(X) \)

(6) \( E(E(X|D)) = E(X) \)

We define an additional function \( D \).