# Lecture 4: Conditional expectation and independence

In elementry probability, conditional probability $P(B|A)$ is defined as $P(B|A) = P(A \cap B)/P(A)$ for events $A$ and $B$ with $P(A) > 0$.
For two random variables, $X$ and $Y$, how do we define $P(X \in B|Y = y)$?

## Definition 1.6

Let $X$ be an integrable random variable on $(\Omega, \mathscr{F}, P)$.

(i) The *conditional expectation* of $X$ given $\mathscr{A}$ (a sub-$\sigma$-field of $\mathscr{F}$), denoted by $E(X|\mathscr{A})$, is the a.s.-unique random variable satisfying the following two conditions:

  (a) $E(X|\mathscr{A})$ is measurable from $(\Omega, \mathscr{A})$ to $(\mathscr{R}, \mathscr{B})$;

  (b) $\int_A E(X|\mathscr{A})dP = \int_A XdP$ for any $A \in \mathscr{A}$.

(ii) The *conditional probability* of $B \in \mathscr{F}$ given $\mathscr{A}$ is defined to be $P(B|\mathscr{A}) = E(I_B|\mathscr{A})$.

(iii) Let $Y$ be measurable from $(\Omega, \mathscr{F}, P)$ to $(\Lambda, \mathscr{G})$.
The conditional expectation of $X$ given $Y$ is defined to be $E(X|Y) = E[X|\sigma(Y)]$.

## Remarks

- The existence of $E(X|\mathscr{A})$ follows from Theorem 1.4.
- $\sigma(Y)$ contains "the information in $Y$"
- $E(X|Y)$ is the "expectation" of $X$ given the information in $Y$
- For a random vector $X$, $E(X|\mathscr{A})$ is defined as the vector of conditional expectations of components of $X$.

## Lemma 1.2

Let $Y$ be measurable from $(\Omega, \mathscr{F})$ to $(\Lambda, \mathscr{G})$ and $Z$ a function from $(\Omega, \mathscr{F})$ to $\mathscr{R}^k$.
Then $Z$ is measurable from $(\Omega, \sigma(Y))$ to $(\mathscr{R}^k, \mathscr{B}^k)$ iff there is a measurable function $h$ from $(\Lambda, \mathscr{G})$ to $(\mathscr{R}^k, \mathscr{B}^k)$ such that $Z = h \circ Y$.

By Lemma 1.2, there is a Borel function $h$ on $(\Lambda, \mathscr{G})$ such that $E(X|Y) = h \circ Y$.
For $y \in \Lambda$, we define $E(X|Y = y) = h(y)$ to be the conditional expectation of $X$ given $Y = y$.
$h(y)$ is a function on $\Lambda$, whereas $h \circ Y = E(X|Y)$ is a function on $\Omega$.

## Example 1.21

Let $X$ be an integrable random variable on $(\Omega, \mathscr{F}, P)$, $A_1, A_2, ...$ be disjoint events on $(\Omega, \mathscr{F}, P)$ such that $\cup A_i = \Omega$ and $P(A_i) > 0$ for all $i$, and let $a_1, a_2, ...$ be distinct real numbers.

Define $Y = a_1 I_{A_1} + a_2 I_{A_2} + \cdots$. We now show that

$$E(X|Y) = \sum_{i=1}^{\infty} \frac{\int_{A_i} XdP}{P(A_i)} I_{A_i}.$$

We need to verify (a) and (b) in Definition 1.6 with $\mathscr{A} = \sigma(Y)$.

Since $\sigma(Y) = \sigma(\{A_1, A_2, ...\})$, it is clear that the function on the right-hand side is measurable on $(\Omega, \sigma(Y))$.

This verifies (a).

To verify (b), we need to show

$$\int_{Y^{-1}(B)} XdP = \int_{Y^{-1}(B)} \left[ \sum_{i=1}^{\infty} \frac{\int_{A_i} XdP}{P(A_i)} I_{A_i} \right] dP.$$

for any $B \in \mathscr{B}$,

## Example 1.21 (continued)

Using the fact that $Y^{-1}(B) = \cup_{i:a_i \in B} A_i$, we obtain

$$\begin{aligned}
\int_{Y^{-1}(B)} XdP &= \sum_{i:a_i \in B} \int_{A_i} XdP \\
&= \sum_{i=1}^{\infty} \frac{\int_{A_i} XdP}{P(A_i)} P\left(A_i \cap Y^{-1}(B)\right) \\
&= \int_{Y^{-1}(B)} \left[ \sum_{i=1}^{\infty} \frac{\int_{A_i} XdP}{P(A_i)} I_{A_i} \right] dP,
\end{aligned}$$

where the last equality follows from Fubini's theorem.
This verifies (b) and thus the result.
Let $h$ be a Borel function on $\mathscr{R}$ satisfying

$$h(a_i) = \int_{A_i} XdP / P(A_i).$$

Then $E(X|Y) = h \circ Y$ and $E(X|Y = y) = h(y)$.

## Proposition 1.9

Let $X$ be a random $n$-vector and $Y$ a random $m$-vector.
Suppose that $(X,Y)$ has a joint p.d.f. $f(x,y)$ w.r.t. $\nu \times \lambda$, where $\nu$ and $\lambda$ are $\sigma$-finite measures on $(\mathscr{R}^n, \mathscr{B}^n)$ and $(\mathscr{R}^m, \mathscr{B}^m)$, respectively.
Let $g(x,y)$ be a Borel function on $\mathscr{R}^{n+m}$ for which $E|g(X,Y)| < \infty$.
Then

$$E[g(X,Y)|Y] = \frac{\int g(x,Y)f(x,Y)d\nu(x)}{\int f(x,Y)d\nu(x)} \quad \text{a.s.}$$

## Proof

Denote the right-hand side by $h(Y)$.
By Fubini's theorem, $h$ is Borel.
Then, by Lemma 1.2, $h(Y)$ is Borel on $(\Omega, \sigma(Y))$.
Also, by Fubini's theorem,

$$f_Y(y) = \int f(x,y)d\nu(x)$$

is the p.d.f. of $Y$ w.r.t. $\lambda$.

## Proof (continued)

For $B \in \mathscr{B}^m$,

$$
\begin{aligned}
\int_{Y^{-1}(B)} h(Y)dP &= \int_B h(y)dP_Y \\
&= \int_B \frac{\int g(x,y)f(x,y)d\nu(x)}{\int f(x,y)d\nu(x)} f_Y(y)d\lambda(y) \\
&= \int_{\mathscr{R}^n \times B} g(x,y)f(x,y)d\nu \times \lambda \\
&= \int_{\mathscr{R}^n \times B} g(x,y)dP_{(X,Y)} \\
&= \int_{Y^{-1}(B)} g(X,Y)dP,
\end{aligned}
$$

where the first and the last equalities follow from Theorem 1.2, the second and the next to last equalities follow from the definition of $h$ and p.d.f.'s, and the third equality follows from Fubini's theorem.

## Conditional p.d.f.

Let $(X, Y)$ be a random vector with a joint p.d.f. $f(x, y)$ w.r.t. $\nu \times \lambda$
The *conditional* p.d.f. of $X$ given $Y = y$ is defined to be

$$f_{X|Y}(x|y) = f(x, y)/f_Y(y)$$

where

$$f_Y(y) = \int f(x, y) d\nu(x)$$

is the marginal p.d.f. of $Y$ w.r.t. $\lambda$.
For each fixed $y$ with $f_Y(y) > 0$, $f_{X|Y}(x|y)$ is a p.d.f. w.r.t. $\nu$.
Then Proposition 1.9 states that

$$E[g(X, Y)|Y] = \int g(x, Y) f_{X|Y}(x|Y) d\nu(x)$$

i.e., the conditional expectation of $g(X, Y)$ given $Y$ is equal to the expectation of $g(X, Y)$ w.r.t. the conditional p.d.f. of $X$ given $Y$.

## Proposition 1.10

Let $X$, $Y$, $X_1, X_2, \dots$ be integrable random variables on $(\Omega, \mathscr{F}, P)$ and $\mathscr{A}$ be a sub-$\sigma$-field of $\mathscr{F}$.

(i) If $X = c$ a.s., $c \in \mathscr{R}$, then $E(X|\mathscr{A}) = c$ a.s.

(ii) If $X \le Y$ a.s., then $E(X|\mathscr{A}) \le E(Y|\mathscr{A})$ a.s.

(iii) If $a, b \in \mathscr{R}$, then $E(aX + bY|\mathscr{A}) = aE(X|\mathscr{A}) + bE(Y|\mathscr{A})$ a.s.

(iv) $E[E(X|\mathscr{A})] = EX$.

(v) $E[E(X|\mathscr{A})|\mathscr{A}_0] = E(X|\mathscr{A}_0) = E[E(X|\mathscr{A}_0)|\mathscr{A}]$ a.s., where $\mathscr{A}_0$ is a sub-$\sigma$-field of $\mathscr{A}$.

(vi) If $\sigma(Y) \subset \mathscr{A}$ and $E|XY| < \infty$, then $E(XY|\mathscr{A}) = YE(X|\mathscr{A})$ a.s.

(vii) If $X$ and $Y$ are independent and $E|g(X,Y)| < \infty$ for a Borel function $g$, then $E[g(X,Y)|Y = y] = E[g(X,y)]$ a.s. $P_Y$.

(viii) If $EX^2 < \infty$, then $[E(X|\mathscr{A})]^2 \le E(X^2|\mathscr{A})$ a.s.

(ix) (Fatou's lemma). If $X_n \ge 0$ for any $n$, then $E\left(\liminf_n X_n \middle| \mathscr{A}\right) \le \liminf_n E(X_n|\mathscr{A})$ a.s.

(x) (Dominated convergence theorem). If $|X_n| \le Y$ for any $n$ and $X_n \to_{a.s.} X$, then $E(X_n|\mathscr{A}) \to_{a.s.} E(X|\mathscr{A})$.

## Example 1.22

Let $X$ be a random variable on $(\Omega, \mathscr{F}, P)$ with $EX^2 < \infty$ and let $Y$ be a measurable function from $(\Omega, \mathscr{F}, P)$ to $(\Lambda, \mathscr{G})$.
One may wish to predict the value of $X$ based on an observed value of $Y$. Let $g(Y)$ be a predictor, i.e.,

$$g \in \aleph = \{\text{all Borel functions } g \text{ with } E[g(Y)]^2 < \infty\}.$$

Each predictor is assessed by the “mean squared prediction error"

$$E[X - g(Y)]^2.$$

We now show that $E(X|Y)$ is the best predictor of $X$ in the sense that

$$E[X - E(X|Y)]^2 = \min_{g \in \aleph} E[X - g(Y)]^2.$$

First, Proposition 1.10(viii) implies $E(X|Y) \in \aleph$.

## Example 1.22 (continued)

Next, for any $g \in \aleph$,

$$\begin{aligned}
E[X-g(Y)]^2 =& E[X-E(X|Y)+E(X|Y)-g(Y)]^2 \\
=& E[X-E(X|Y)]^2+E[E(X|Y)-g(Y)]^2 \\
& +2E\{[X-E(X|Y)][E(X|Y)-g(Y)]\} \\
=& E[X-E(X|Y)]^2+E[E(X|Y)-g(Y)]^2 \\
& +2E\big\{E\{[X-E(X|Y)][E(X|Y)-g(Y)]|Y\}\big\} \\
=& E[X-E(X|Y)]^2+E[E(X|Y)-g(Y)]^2 \\
& +2E\{[E(X|Y)-g(Y)]E[X-E(X|Y)|Y]\} \\
=& E[X-E(X|Y)]^2+E[E(X|Y)-g(Y)]^2 \\
\geq& E[X-E(X|Y)]^2,
\end{aligned}$$

where the third equality follows from Proposition 1.10(iv), the fourth equality follows from Proposition 1.10(vi), and the last equality follows from Proposition 1.10(i), (iii), and (vi).

## Definition 1.7 (Independence).

Let $(\Omega, \mathscr{F}, P)$ be a probability space.

(i) Let $\mathscr{C}$ be a collection of subsets in $\mathscr{F}$.
Events in $\mathscr{C}$ are said to be *independent* iff for any positive integer $n$ and distinct events $A_1$,...,$A_n$ in $\mathscr{C}$,

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2)\cdots P(A_n).$$

(ii) Collections $\mathscr{C}_i \subset \mathscr{F}$, $i \in \mathscr{I}$ (an index set that can be uncountable), are said to be independent iff events in any collection of the form $\{A_i \in \mathscr{C}_i : i \in \mathscr{I}\}$ are independent.

(iii) Random elements $X_i$, $i \in \mathscr{I}$, are said to be independent iff $\sigma(X_i)$, $i \in \mathscr{I}$, are independent.

## Lemma 1.3 (a useful result for checking the independence of $\sigma$-fields)

Let $\mathscr{C}_i$, $i \in \mathscr{I}$, be independent collections of events.
If each $\mathscr{C}_i$ is a $\pi$-system ($A \in \mathscr{C}_i$ and $B \in \mathscr{C}_i$ implies $A \cap B \in \mathscr{C}_i$), then $\sigma(\mathscr{C}_i)$, $i \in \mathscr{I}$, are independent.

## Facts

- Random variables $X_i$, $i = 1, ..., k$, are independent according to Definition 1.7 iff

$$F_{(X_1,...,X_k)}(x_1,...,x_k) = F_{X_1}(x_1)\cdots F_{X_k}(x_k), \quad (x_1,...,x_k) \in \mathscr{R}^k$$

  Take $\mathscr{C}_i = \{(a,b] : a \in \mathscr{R}, b \in \mathscr{R}\}$, $i = 1, ..., k$
- If $X$ and $Y$ are independent random vectors, then so are $g(X)$ and $h(Y)$ for Borel functions $g$ and $h$.
- Two events $A$ and $B$ are independent iff $P(B|A) = P(B)$, which means that $A$ provides no information about the probability of the occurrence of $B$.

## Proposition 1.11

Let $X$ be a random variable with $E|X| < \infty$ and let $Y_i$ be random $k_i$-vectors, $i = 1, 2$.
Suppose that $(X, Y_1)$ and $Y_2$ are independent.
Then

$$E[X|(Y_1, Y_2)] = E(X|Y_1) \text{ a.s.}$$

## Proof

First, $E(X|Y_1)$ is Borel on $(\Omega, \sigma(Y_1, Y_2))$, since $\sigma(Y_1) \subset \sigma(Y_1, Y_2)$.
Next, we need to show that for any Borel set $B \in \mathscr{B}^{k_1+k_2}$,

$$\int_{(Y_1,Y_2)^{-1}(B)} X dP = \int_{(Y_1,Y_2)^{-1}(B)} E(X|Y_1) dP.$$

If $B = B_1 \times B_2$, where $B_i \in \mathscr{B}^{k_i}$, then

$$(Y_1, Y_2)^{-1}(B) = Y_1^{-1}(B_1) \cap Y_2^{-1}(B_2)$$

and

$$\begin{aligned}
\int_{Y_1^{-1}(B_1)\cap Y_2^{-1}(B_2)} E(X|Y_1) dP &= \int I_{Y_1^{-1}(B_1)} I_{Y_2^{-1}(B_2)} E(X|Y_1) dP \\
&= \int I_{Y_1^{-1}(B_1)} E(X|Y_1) dP \int I_{Y_2^{-1}(B_2)} dP \\
&= \int I_{Y_1^{-1}(B_1)} X dP \int I_{Y_2^{-1}(B_2)} dP \\
&= \int I_{Y_1^{-1}(B_1)} I_{Y_2^{-1}(B_2)} X dP \\
&= \int_{Y_1^{-1}(B_1)\cap Y_2^{-1}(B_2)} X dP,
\end{aligned}$$

where the second and the next to last equalities follow the independence of $(X, Y_1)$ and $Y_2$, and the third equality follows from the fact that $E(X|Y_1)$ is the conditional expectation of $X$ given $Y_1$.

This shows that the result for $B = B_1 \times B_2$.

Note that $\mathscr{B}^{k_1} \times \mathscr{B}^{k_2}$ is a $\pi$-system.

We can show that the following collection is a $\lambda$-system:

$$\mathscr{H} = \left\{ B \subset \mathscr{R}^{k_1+k_2} : \int_{(Y_1,Y_2)^{-1}(B)} X dP = \int_{(Y_1,Y_2)^{-1}(B)} E(X|Y_1) dP \right\}$$

Since we have already shown that $\mathscr{B}^{k_1} \times \mathscr{B}^{k_2} \subset \mathscr{H}$, $\mathscr{B}^{k_1+k_2} = \sigma(\mathscr{B}^{k_1} \times \mathscr{B}^{k_2}) \subset \mathscr{H}$ and thus the result follows.

## Remark

The result in Proposition 1.11 still holds if $X$ is replaced by $h(X)$ for any Borel $h$ and, hence,

$$P(A|Y_1, Y_2) = P(A|Y_1) \text{ a.s. for any } A \in \sigma(X),$$

if $(X, Y_1)$ and $Y_2$ are independent.

## Conditional independence

Let $X$, $Y$, and $Z$ be random vectors.
We say that given $Z$, $X$ and $Y$ are *conditionally independent* iff

$$P(A|X,Z) = P(A|Z) \text{ a.s. for any } A \in \sigma(Y).$$

Proposition 1.11 can be stated as: if $Y_2$ and $(X, Y_1)$ are independent, then given $Y_1$, $X$ and $Y_2$ are conditionally independent.

## Discussion

- Conditional independence is a very important concept for statistics.
- For example, if $X$ and $Z$ are covariates associated with a response $Y$, and if given $Z$, $X$ and $Y$ are conditionally independent, then when we have $Z$, we do not need $X$ to study the relationship between $Y$ and covariates.
- The dimension of the covariate vector is reduced without losing information (sufficient dimension reduction).
- Although $X$ may be unconditionally dependent of $Y$, it is related with $Y$ through $Z$.