In elementary probability, conditional probability $P(B|A)$ is defined as

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

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 $(\Omega, \mathcal{F}, P)$.

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

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

(b) $\int_{\mathcal{A}} E(X|\mathcal{A}) dP = \int_{\mathcal{A}} X dP$ for any $A \in \mathcal{A}$.

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

(iii) Let $Y$ be measurable from $(\Omega, \mathcal{F}, P)$ to $(\Lambda, \mathcal{G})$.

The conditional expectation of $X$ given $Y$ is defined to be $E(X|Y) = E[X|\sigma(Y)]$. 

UW-Madison (Statistics) 2018 1 / 15
Remarks

- The existence of $E(X|\mathcal{A})$ 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 $X$, $E(X|\mathcal{A})$ is defined as the vector of conditional expectations of components of $X$.

Lemma 1.2

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

By Lemma 1.2, there is a Borel function $h$ on $(\Lambda, \mathcal{G})$ 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 $(\Omega, \mathcal{F}, P)$, $A_1, A_2, \ldots$ be disjoint events on $(\Omega, \mathcal{F}, P)$ such that $\bigcup A_i = \Omega$ and $P(A_i) > 0$ for all $i$, and let $a_1, a_2, \ldots$ be distinct real numbers. Define $Y = a_1 I_{A_1} + a_2 I_{A_2} + \cdots$. We now show that

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

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

Since $\sigma(Y) = \sigma(\{A_1, A_2, \ldots\})$, 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

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

for any $B \in \mathcal{B}$,
Example 1.21 (continued)

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

\[
\int_{Y^{-1}(B)} X dP = \sum_{i: a_i \in B} \int_{A_i} X dP
\]

\[
= \sum_{i=1}^{\infty} \frac{\int_{A_i} X dP}{P(A_i)} P \left( A_i \cap Y^{-1}(B) \right)
\]

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

where the last equality follows from Fubini’s theorem.

This verifies (b) and thus the result.

Let \( h \) be a Borel function on \( \mathbb{R} \) satisfying

\[
h(a_i) = \int_{A_i} X dP / P(A_i).
\]

Then \( E(X \mid Y) = h \circ Y \) and \( E(X \mid Y = y) = h(y) \).
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 $(\mathbb{R}^n, \mathcal{B}^n)$ and $(\mathbb{R}^m, \mathcal{B}^m)$, respectively. Let $g(x, y)$ be a Borel function on $\mathbb{R}^{n+m}$ for which $E|g(X, Y)| < \infty$. Then

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

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 $B \in \mathcal{B}^m$, 

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

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

\[
f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}
\]

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

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

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

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

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

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

(vi) If $\sigma(Y) \subset \mathcal{A}$ and $E|XY| < \infty$, then $E(XY|\mathcal{A}) = YE(X|\mathcal{A})$ 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 $[E(X|\mathcal{A})]^2 \leq E(X^2|\mathcal{A})$ a.s.

(ix) (Fatou’s lemma). If $X_n \geq 0$ for any $n$, then $E \left( \liminf_n X_n \big| \mathcal{A} \right) \leq \liminf_n E(X_n|\mathcal{A})$ a.s.

(x) (Dominated convergence theorem). If $|X_n| \leq Y$ for any $n$ and $X_n \to_{a.s.} X$, then $E(X_n|\mathcal{A}) \to_{a.s.} E(X|\mathcal{A})$. 
Example 1.22

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

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

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

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

First, Proposition 1.10(viii) implies $E(X|Y) \in \mathcal{X}$. 
Example 1.22 (continued)

Next, for any \( g \in \mathfrak{X} \),

\[
E[X - g(Y)]^2 = E[X - E(X|Y) + E(X|Y) - g(Y)]^2 \\
= E[X - E(X|Y)]^2 + E[E(X|Y) - g(Y)]^2 \\
+ 2E\{[X - E(X|Y)][E(X|Y) - g(Y)]\} \\
= E[X - E(X|Y)]^2 + E[E(X|Y) - g(Y)]^2 \\
+ 2E\{E\{[X - E(X|Y)][E(X|Y) - g(Y)]|Y\}\} \\
= E[X - E(X|Y)]^2 + E[E(X|Y) - g(Y)]^2 \\
+ 2E\{[E(X|Y) - g(Y)]E[X - E(X|Y)|Y]\} \\
= E[X - E(X|Y)]^2 + E[E(X|Y) - g(Y)]^2 \\
\geq E[X - E(X|Y)]^2,
\]

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 \((\Omega, \mathcal{F}, P)\) be a probability space.

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

\[
P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2) \cdots P(A_n).
\]

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

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

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

Let \(C_i, i \in I\), be independent collections of events. If each \(C_i\) is a \(\pi\)-system (\(A \in C_i\) and \(B \in C_i\) implies \(A \cap B \in C_i\)), then \(\sigma(C_i), i \in I\), are independent.
Facts

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

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

Take $C_i = \{(a, b) : a \in \mathbb{R}, b \in \mathbb{R}\}$, $i = 1, \ldots, k$

- 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 $B \in \mathcal{B}^{k_1+k_2}$,

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

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

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

and

$$\int_{Y_1^{-1}(B_1) \cap Y_2^{-1}(B_2)} E(X|Y_1)dP = \int I_{Y_1^{-1}(B_1)} I_{Y_2^{-1}(B_2)} E(X|Y_1)dP = \int I_{Y_1^{-1}(B_1)} E(X|Y_1)dP \int I_{Y_2^{-1}(B_2)} dP = \int I_{Y_1^{-1}(B_1)} XdP \int I_{Y_2^{-1}(B_2)} dP = \int I_{Y_1^{-1}(B_1)} I_{Y_2^{-1}(B_2)} XdP = \int_{Y_1^{-1}(B_1) \cap Y_2^{-1}(B_2)} XdP,$$
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 $\mathcal{B}_{k_1}^k \times \mathcal{B}_{k_2}^k$ is a $\pi$-system.

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

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

Since we have already shown that $\mathcal{B}_{k_1}^k \times \mathcal{B}_{k_2}^k \subset \mathcal{H}$, $\mathcal{B}_{k_1+k_2} = \sigma(\mathcal{B}_{k_1}^k \times \mathcal{B}_{k_2}^k) \subset \mathcal{H}$ 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 $Z$. 