Lecture 24: Conditional Expectation

1.) Definition and Properties

**Definition:** Recall that, if \( X \) and \( Y \) are jointly discrete random variables, then the conditional probability mass function of \( X \) given \( Y = y \) is defined to be
\[
p_{X|Y}(x|y) = \frac{P(X = x|Y = y)}{p_Y(y)},
\]
provided \( p_Y(y) > 0 \). In this case, we can define the **conditional expectation** of \( X \) given \( Y = y \) by
\[
E[X|Y = y] = \sum_x x P(X = x|Y = y) = \sum_x p_{X|Y}(x|y) \cdot x.
\]

Similarly, if \( X \) and \( Y \) are jointly continuous with conditional density function \( p_{X|Y}(x|y) \), then the conditional expectation of \( X \) given \( Y = y \) is defined by
\[
E[X|Y = y] = \int_{-\infty}^{\infty} x \cdot p_{X|Y}(x|y) dx.
\]

**Example:** Suppose that \( X \) and \( Z \) are independent Poisson RVs, both with parameter \( \lambda \), and let \( Y = X + Z \). Find the conditional expectation of \( X \) given that \( Y = n \).

**Solution:** It follows from a result in lecture (21) that the conditional distribution of \( X \) given that \( Y = n \) is binomial with parameters \((n, 1/2)\):
\[
p_{X|Y}(k|n) = \binom{n}{k} \left( \frac{1}{2} \right)^n,
\]
for \( k = 0, \ldots, n \). Consequently,
\[
E[X|Y = n] = \sum_{k=0}^{n} k \binom{n}{k} \left( \frac{1}{2} \right)^n = \frac{n}{2}.
\]

We can also solve this problem by observing that
\[
n = E[Y|Y = n] = E[X|Y = n] + E[Z|Y = n] = 2E[X|Y = n],
\]
using the fact that \( E[X|Y = n] = E[Z|Y = n] \) since \( X \) and \( Y \) have the same conditional distribution given \( Z \). \( \square \)

**Remark:** In general, all of the results that hold for expectations also hold for conditional expectations, including
\[
E[g(X)|Y = y] = \begin{cases} 
\sum_x g(x)p_{X|Y}(x|y) & \text{discrete case} \\
\int_{-\infty}^{\infty} g(x)p_{X|Y}(x|y) dx & \text{continuous case},
\end{cases}
\]
and
\[
E \left[ \sum_{i=1}^{n} X_i | Y = y \right] = \sum_{i=1}^{n} E[X_i | Y = y].
\]

2.) Computing Expectations by Conditioning

**Definition:** The conditional expectation of \( X \) given \( Y \) is the random variable
\[
E[X | Y] = H(Y),
\]
where \( H(y) \) is the (deterministic) function defined by the formula
\[
H(y) = E[X | Y = y].
\]

The next proposition provides us with a powerful tool for calculating expectations by conditioning on a second random variable.

**Proposition 5.1:** For any two random variables \( X \) and \( Y \), we have
\[
E[X] = E[E[X | Y]],
\]
provided that both \( E[X] \) and \( E[X | Y] \) exist. In particular, if \( Y \) is discrete, then
\[
E[X] = \sum_y E[X | Y = y]P\{Y = y\},
\]
while if \( Y \) is continuous with density \( p_Y(y) \), then
\[
E[X] = \int_{-\infty}^{\infty} E[X | Y = y]p_Y(y)dy.
\]

**Proof:** We will assume that \( X \) and \( Y \) are both discrete. Then
\[
\sum_y E[X | Y = y]P\{Y = y\} = \sum_y \sum_x xP\{X = x | Y = y\}P\{Y = y\}
= \sum_y \sum_x \frac{xP\{X = x, Y = y\}}{P\{Y = y\}}P\{Y = y\}
= \sum_y \sum_x xP\{X = x, Y = y\}
= \sum_x x \sum_y P\{X = x, Y = y\}
= \sum_x xP\{X = x\}
= E[X].
\]
\]
Example (Ross, 5d): Wald’s Identity. Let $N$ be a RV with values in the natural numbers and finite mean $\mathbb{E}[N] < \infty$, and let $X_1, X_2, \cdots$ be a sequence of independent, identically-distributed random variables that are independent of $N$ and have finite mean $\mathbb{E}[X]$. Our goal is to calculate the expected value of the random sum $X_1 + \cdots + X_N$, which we can do by conditioning on $N$:

$$\mathbb{E} \left[ \sum_{i=1}^{N} X_i \right] = \mathbb{E} \left[ \mathbb{E} \left[ \sum_{i=1}^{N} X_i | N \right] \right].$$

Notice that

$$\mathbb{E} \left[ \sum_{i=1}^{N} X_i | N = n \right] = \mathbb{E} \left[ \sum_{i=1}^{n} X_i | N = n \right] = \mathbb{E} \left[ \sum_{i=1}^{n} X_i \right] \quad \text{since } X_i \text{ and } N \text{ are independent}$$

$$= n \mathbb{E}[X],$$

which implies that

$$\mathbb{E} \left[ \sum_{i=1}^{N} X_i | N \right] = N \mathbb{E}[X]$$

and consequently

$$\mathbb{E} \left[ \sum_{i=1}^{N} X_i \right] = \mathbb{E}[N \cdot \mathbb{E}[X]] = \mathbb{E}[N] \cdot \mathbb{E}[X].$$

□

Example: A discrete random variable $X$ is said to have a **compound Poisson distribution** if there exists a Poisson RV $N$ with parameter $\lambda > 0$ and a set of IID random variables $Y_1, Y_2, \cdots$, all independent of $N$, such that

$$X = \sum_{i=1}^{N} Y_i.$$

If $\mathbb{E}[Y_i] = \mu_y < \infty$, then by Wald’s identity we have

$$\mathbb{E}[X] = \lambda \cdot \mu_y.$$

4.) Computing Probabilities by Conditioning

Conditioning is also sometimes helpful when calculating probabilities. Suppose that $E$ is some event and let $X$ be the indicator function of $E$:

$$X = \begin{cases} 
1 & \text{if } E \text{ occurs} \\
0 & \text{otherwise.}
\end{cases}$$
Then, for any random variable $Y$ defined on the same probability space as $X$,
\[
\mathbb{E}[X] = \mathbb{P}(E)
\]
\[
\mathbb{E}[X|Y = y] = \mathbb{P}(E|Y = y).
\]
Furthermore, by using Proposition 5.1, we have
\[
\mathbb{P}(E) = \sum_y \mathbb{P}(E|Y = y)\mathbb{P}(Y = y) \quad \text{if } Y \text{ is discrete}
\]
\[
= \int_{-\infty}^{\infty} \mathbb{P}(E|Y = y)p_Y(y)dy \quad \text{if } Y \text{ is continuous}.
\]

**Example (Ross, 5I):** Let $U$ be a uniform RV on $(0,1)$, and suppose that the conditional distribution of $X$, given $U = p$, is binomial with parameters $(n, p)$. Then the probability mass function of $X$ can be calculated by conditioning on $p$:
\[
\mathbb{P}\{X = k\} = \mathbb{E}[\mathbb{P}\{X = k|U = p\}]\]
\[
= \int_0^1 \mathbb{P}\{X = k|U = p\}dp
\]
\[
= \int_0^1 \left(\frac{n}{k}\right)p^k(1-p)^{n-k}dp
\]
\[
= \frac{n!}{k!(n-k)!} \int_0^1 p^k(1-p)^{n-k}dp.
\]

Using the Beta function, it can be shown that the value of the integral in this last line is
\[
\int_0^1 p^k(1-p)^{n-k}dp = \beta(k+1, n-k+1) = \frac{k!(n-k)!}{(n+1)!},
\]
so upon substituting this expression back into the preceding series of equations, we obtain:
\[
\mathbb{P}\{X = k\} = \frac{1}{n+1},
\]
i.e., $X$ is uniformly distributed on the set $\{0, 1, \cdots, n\}$. $X$ is an example of a **mixture model**: the distribution of $X$ is equal to the mixture of a family of probability distributions with respect to another distribution called the mixing distribution. These play an important role in statistics.

5.) **Conditional Variance**

**Definition:** If $X$ and $Y$ are random variables defined on the same probability space, then the **conditional variance** of $X$ given that $Y = y$ is the quantity
\[
\text{Var}(X|Y = y) \equiv \mathbb{E}[(X - \mathbb{E}[X|Y = y])^2|Y = y].
\]
Similarly, the conditional variance of $X$ given $Y$ is the random variable
\[
\text{Var}(X|Y) \equiv \mathbb{E}[(X - \mathbb{E}[X|Y])^2|Y].
\]
As the following proposition shows, in general, the expected conditional variance of a random variable is less than its variance.

**Proposition 2.1: Law of Total Variance**

\[ Var(X) = \mathbb{E}[Var(X|Y)] + Var(\mathbb{E}[X|Y]). \]

**Proof:** First observe that

\[ Var(X|Y) = \mathbb{E}[X^2|Y] - (\mathbb{E}[X|Y])^2, \]

and so

\[
\mathbb{E}[Var(X|Y)] = \mathbb{E}[\mathbb{E}[X^2|Y]] - \mathbb{E}[(\mathbb{E}[X|Y])^2] = \mathbb{E}[X^2] - \mathbb{E}[(\mathbb{E}[X|Y])^2].
\]

Also, because \( \mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X] \), the variance of the conditional expectation of \( X \) given \( Y \) can be written as

\[ Var(\mathbb{E}[X|Y]) = \mathbb{E}[(\mathbb{E}[X|Y])^2] - (\mathbb{E}[X])^2. \]

The law of total variance follows upon adding these two equations.

\[ \square \]

**Example (Ross, 5p):** Let \( N \) and \( X_1, X_2, \cdots \) be as in Example (5d), with the added assumption that \( Var(N) < \infty \) and \( Var(X) < \infty \). Then

\[ \mathbb{E} \left[ \sum_{i=1}^{N} X_i | N \right] = N \cdot \mathbb{E}[X] \]

\[ Var \left( \sum_{i=1}^{N} X_i | N \right) = N \cdot Var(X), \]

and so the law of total variance implies that

\[ Var \left( \sum_{i=1}^{N} X_i \right) = \mathbb{E}[N] \cdot Var(X) + (\mathbb{E}[X])^2 \cdot Var(N). \]