18. Conditional expectation and variance

We define the conditional expectation of $Y$ given $X$, written $E(Y \mid X)$, by

$$E(Y \mid X = x) := \begin{cases} \sum_y y f_{Y\mid X}(y \mid x) & \text{if } Y \text{ is discrete} \\ \int y f_{Y\mid X}(y \mid x) \, dy & \text{if } Y \text{ is continuous.} \end{cases}$$

We define the conditional variance of $Y$ given $X$, denoted $\text{Var}(Y \mid X)$, analogously.

**Example 18.1.** Suppose $X, Y$ are jointly normal with $\mathbb{E}X = \mathbb{E}Y = 0$, $\text{Var}(X) = \text{Var}(Y) = 1$, and $\text{Cov}(X, Y) = \rho$. Then $Y \mid X = x \sim N(\rho x, 1 - \rho^2)$, and

$$E(Y \mid X) = \rho x \quad \text{and} \quad \text{Var}(Y \mid X) = 1 - \rho^2.$$ 

Conditional expectation has all the usual properties of expectation since it is essentially the expectation you would compute for the reduced sample space $\{\omega \in \Omega : X(\omega) = x\}$; however, you might want to keep in mind (and come to terms with) the fact that this reduced sample space is in fact random, it depends on the value of the random variable $X$. In that sense, $E(Y \mid X)$ and $\text{Var}(Y \mid X)$ are, themselves, random variables.

We also have:

$$E(g(X)h(Y) \mid X = x) = g(x)E(h(Y) \mid X = x)$$

$$E(g(X) \mid X = x) = g(x)$$

$$E(Y \mid X = x) = E(Y) \quad \text{if } X \text{ and } Y \text{ are independent.}$$

18.1. Conditional mean and variance of $Y$ given $X$. For each $x$, let $\psi(x) := E(Y \mid X = x)$. The random variable $\psi(X)$ is the conditional mean of $Y$ given $X$, denoted $E(Y \mid X)$. The conditional mean satisfies the tower property of conditional expectation:

$$\mathbb{E}Y = \mathbb{E}(\mathbb{E}(Y \mid X)),$$

which coincides with the law of cases for expectation. To define conditional variance $\text{Var}(Y \mid X) := \psi(X)$, where $\psi(x) := \text{Var}(Y \mid X = x)$, we need $\mathbb{E}|Y| < \infty$.

**Example 18.2.** Let $X$ and $Y$ be marginally standard Normal random variables with $\text{Cov}(X, Y) = \rho$. We have already seen that the conditional law of $Y$ given $X = x$ is $N(\rho x, 1 - \rho^2)$. We have

- $E(Y \mid X = x) = \rho x$ implies $E(Y \mid X) = \rho X$.
- $\text{Var}(Y \mid X = x) = 1 - \rho^2$ implies $\text{Var}(Y \mid X) = 1 - \rho^2$.
- $E(Y \mid X)$ has mean $\rho \mathbb{E}X = 0 = \mathbb{E}Y$.
- $\text{Var}(Y \mid X)$ has mean $1 - \rho^2$.

We observe

$$\text{Var}(E(Y \mid X)) + \mathbb{E}\text{Var}(Y \mid X) = \rho^2 + (1 - \rho^2) = 1 = \text{Var}(Y).$$

18.2. Mean and variance of a random sum. Consider the random sum

$$S_N := \sum_{i=1}^N X_i = X_1 + \cdots + X_N,$$

where $X_1, X_2, \ldots$ are independent, identically distributed and independent of $N$, $\mathbb{E}X_1 = \mu$, $\text{Var}(X_1) = \sigma^2$, and $N$ is a non-negative integer-valued random variable. We have already computed $\mathbb{E}S_N$ and $\text{Var}(S_N)$ during our discussion on generating functions. Alternatively, we can compute these using properties of conditional expectation and variance.
For \( E_{S_N} \), we observe

\[
E_{S_N} = \mathbb{E}(S_N | N = n) \\
= \mathbb{E}(S_n | N = n) \\
= \mathbb{E}(N_{i\in EX}) \\
= \mathbb{E}N_{i\in EX}.
\]

For \( \text{Var}(S_N) \), we have

\[
\text{Var}(S_N | N = n) = \text{Var}(S_n) = n \text{Var}(X), \quad \text{which implies} \quad \text{Var}(S_N | N) = N \text{Var}(X).
\]

From above, we have \( E_{S_N} = \mathbb{E}N_{i\in EX} \) and so \( \text{Var}(E(S_N | N)) = \text{Var}(N_{i\in EX}) \). We conclude

\[
\text{Var}(S_N) = [\mathbb{E}X]^2 \text{Var}(N) + \mathbb{E}N \text{Var}(X).
\]

Note that (i) \( \mathbb{E}(E(Y | X)) = \mathbb{E}Y \) and (ii) \( \text{Var}(Y) = \text{Var}(E(Y | X)) + \text{Var}(Y | X) \). To see (ii):

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

**Example 18.3.** Roll a fair 10-sided die. If we roll a 1, 2, \ldots, 7, we win that many dollars and start over (roll again). If we roll 8, 9, 10, we win that many dollars and the game stops. What is the expectation and variance of the number of rolls and the amount of money won?

Consider a coin with \( P(\text{Heads}) = 3/10 \). Each time it lands \text{Tails}, we get a random number \( U_i \sim d-\text{Unif}[1, \ldots, 7] \) of dollars. If it lands on \text{Heads}, we get \( V \sim d-\text{Unif}[8, 9, 10] \) and quit. Let \( N \) be the number of tails flipped before the first head appears. Then \( N \sim \text{Geometric}(3/10) \). The number of dollars won is \( M := V + S_N \), where \( S_N = U_1 + \cdots + U_N \), \( N \sim \text{Geo}(3/10) \) and \( P(U_i = j) = 1/7 \), \( j = 1, \ldots, 7 \). Hence,

\[
\mathbb{E}N = q/p = 7/3, \quad \mathbb{E}U_1 = 4, \quad \mathbb{E}V = 9 \quad \text{and} \quad Var(N) = q/p^2 = 70/9, \quad Var(U_1) = \frac{7^2 - 1}{12} = 4, \quad Var(U_2) = 2/3.
\]

It follows that

\[
\mathbb{E}M = \mathbb{E}(S_N + V) = \mathbb{E}N\mathbb{E}U_1 + \mathbb{E}V = \frac{7}{3} \times 4 + 9 = 55/3;
\]

\[
\text{Var}(M) = \text{Var}(S_N) + \text{Var}(V) \\
= (\mathbb{E}U_1)^2 \text{Var}(N) + \text{Var}(U_1)\mathbb{E}N + \text{Var}(V) \\
= 4^2 \times 70/9 + 4 \times 7/3 + 2/3 \\
= 134 \times 4/9.
\]

18.3. **Best prediction.** Recall that the constant that minimizes \( \mathbb{E}(Y - c)^2 \) is \( c = \mathbb{E}Y \), provided \( \mathbb{E}Y^2 < \infty \).

Now, suppose \( X, Y \) are random variables with joint probability mass function \( f_{X,Y} \). Then what function \( g(X) \) predicts \( Y \) in terms of minimizing

\[
\text{MSE} = \mathbb{E}(Y - g(X))^2?
\]
We have
\[
\text{MSE} = \mathbb{E}(E((Y - g(X))^2 | X))
= \sum_x E((Y - g(X))^2 | X = x)f_X(x)
= \sum_x E((Y - g(x))^2 | X = x)f_X(x).
\]

For each \(x\), the \(n\)th summand is minimized by \(g(x) = E(Y|X = x)\); hence, the best predictor is \(g(X) = E(Y|X)\), which has MSE
\[
\mathbb{E}(Y - E(Y|X))^2 = \mathbb{E}(E(Y - E(Y|X))^2|X) = \mathbb{E}\text{Var}(Y|X) = \text{Var}(Y) - \text{Var}(E(Y|X)).
\]

18.3.1. **Best linear prediction.** A predictor \(g(X)\) of the form \(a + bX\) is a linear predictor. Suppose \(X, Y\) are mean 0, variance 1 and have \(\text{Cov}(X, Y) = \rho\). What pair \((a, b)\) minimizes \(\mathbb{E}(Y - (a + bX))^2\)? To answer this, notice that
\[
\text{Cov}(Y - \rho X, X) = \text{Cov}(Y, X) - \text{Cov}(\rho X, X) = 0.
\]

Therefore, for \(R = Y - (a + bX) = -a + Y - \rho X + (\rho - b)X\), we have
\[
\mathbb{E}R = -a
\]
\[
\text{Var}(R) = \text{Var}(Y - \rho X) + (\rho - b)^2 \text{Var}(X)
\]
\[
\text{MSE} = \mathbb{E}R^2 = a^2 + (\rho - b)^2 + \text{Var}(Y - \rho X).
\]

Take \(a = 0\) and \(b = \rho\). Then \(\text{MSE}(\rho X) = \text{Var}(Y - \rho X) = 1 - \rho^2\). As an immediate corollary, we observe that \(|\rho| \leq 1\). In general, for arbitrary means \(\mu_X, \mu_Y\), variance \(\sigma_X^2, \sigma_Y^2\), and covariance \(\sigma_{XY}\), the best linear predictor of \(Y\) given \(X\) is
\[
\mu_Y + \rho \frac{\sigma_Y}{\sigma_X} (X - \mu_X),
\]
where \(\rho = \sigma_{XY}/(\sigma_X\sigma_Y)\), with MSE = \(\sigma_Y^2(1 - \rho^2)\).