Chapter 7

Conditional Expectation

This chapter is devoted to a very simple one-line equation. However, this simple equation contains the essence of stochastic processes. We will consider the two-variable case in detail, and consider the generalisation towards the end of the chapter.

7.1 Expectations

Suppose we have a pair of discrete random variable $\{(X, Y)\}$, with an associated joint probability tabulated mass, $f_{XY}(x, y)$, as tabulated below.

		X		
		1	2	3
	-1	0.1	0.1	0.0
Y	0	0.2	0.0	0.3
	2	0.1	0.2	0.0

(7.1)

First, it is prudent to check that the tabulated results satisfy one of the basic requirements of a probability mass

$$\sum_{i,j} f_{X,Y}(x_i, y_j) = 1 \qquad . \tag{7.2}$$

Then to calculate the *marginal masses*, as usual we sum along either the rows or the columns, respectively. This gives us:

$$f_Y(-1) = 0.2 \qquad f_Y 0(0) = 0.5 \qquad f_Y(2) = 0.3 \tag{7.3}$$

While summing the columns we have:

$$f_X(1) = 0.4 \qquad f_X(2) = 0.3 \qquad f_X(3) = 0.3 \qquad . \tag{7.4}$$

Then the calculation of expectations is straightforward:

$$\mathbb{E}(X) = \sum_i x_i f_X(x_i) = 1(0.4) + 2(0.3) + 3(0.3) = 0.4 + 0.6 + 0.9 = 1.9 \tag{7.5}$$

$$\mathbb{E}(Y) = \sum_j y_j f_Y(y_j) = (-1)(0.2) + (0)(0.5) + 2(0.3) = -0.2 + 0.0 + 0.6 = 0.4 \tag{7.6}$$

so that, $\mathbb{E}(X)\,\mathbb{E}(Y) = 0.76$

After some calculation, summing over the entire table:

$$\mathbb{E}(XY) = \sum_{ij} x_i y_j f_{XY}(x_i, y_j) = 0.70 \qquad . \tag{7.7}$$

Since $\mathbb{E}(XY) = 0.7 \neq \mathbb{E}(X)\,\mathbb{E}(Y)$, then clearly X, Y *are* correlated.

We define a *conditional mass probability*:

$$f_{X|Y}(x,y) \equiv P(X=x|Y=y) = \frac{P(X=x,Y=y)}{P(Y=y)} \tag{7.8}$$

That is:

$$f_{X|Y}(x,y) \equiv \frac{f_{XY}(x,y)}{f_Y(y)} \tag{7.9}$$

and similarly:

$$f_{Y|X}(y,x) \equiv \frac{f_{XY}(x,y)}{f_X(x)} \tag{7.10}$$

with the corresponding "multiplication rule"

$$f_{XY}(x,y) = f_{Y|X}(y,x)f_X(x) = f_{X|Y}(x,y)f_Y(y). \tag{7.11}$$

This *conditional probability* obeys all the axioms of conventional (unconditional) probability. For example, $P(Y|X)$, where $P(X) > 0$, satisfies the following relations:

$$P(\emptyset|X) = 0 \qquad \text{and} \qquad P(\Omega|X) = 1 \quad .$$

and if $Y_1 \cap Y_2 = \emptyset$, then:

$$P(Y_1 \cup Y_2|X) = P(Y_1|X) + P(Y_2|X) \quad .$$

7.2 Conditional Expectations

Similarly, one defines *conditional expectations* in a manner analogous to the conventional definition. The *conditional expectation* for Y, *given* that X has a prescribed value, is defined as follows:

$$\mathbb{E}\,(Y|X=x) \equiv \sum_j y_j P(Y=y_j|X=x) = \sum_j y_j f_{Y|X}(x,y_j) \qquad . \tag{7.12}$$

Of course the expectation value depends on X, which acts as a parameter (fixed variable). Naturally, this extends to the Y variable as well:

$$\mathbb{E}\,(X|Y=y) \equiv \sum_i x_i P(X=x_i|Y=y) = \sum_i x_i f_{X|Y}(x_i,y) \qquad . \tag{7.13}$$

In this case, $\mathbb{E}\,(X|Y=y)$ depends on Y.

Let us proceed to derive the *conditional expectation theorem*. First denote the conditional expectation of Y given X as follows:

$$\varphi_X(x) \equiv \mathbb{E}\,(Y|X=x) \tag{7.14}$$

Technically, this is called the *regression function* of X on Y, and has a very important role in statistics. Similarly, the function

$$\psi_Y(y) \equiv \mathbb{E}\,(X|Y=y) \qquad . \tag{7.15}$$

is termed the *regression function* of Y on X.

The expectation of the regression function is:

$$\mathbb{E}\,(\varphi_X(X)) \equiv \sum_i \varphi_X(x_i) f_X(x_i) \qquad . \tag{7.16}$$

This expression can be expanded, and then contracted as follows:

$$\begin{aligned}
\mathbb{E}\left(\mathbb{E}\left(Y|X\right)\right) &= \sum_i \varphi_X(x_i) f_X(x_i) && (7.17)\\
&= \sum_i \left\{ \sum_j y_j \frac{f_{XY}(x_i, y_j)}{f_X(x_i)} \right\} f_X(x_i) && (7.18)\\
&= \sum_j y_j \sum_i f_{XY}(x_i, y_j) && (7.19)\\
&= \sum_j y_j f_Y(y_j) && (7.20)\\
&= \mathbb{E}\left(Y\right) && (7.21)
\end{aligned}$$

Thus we have the *conditional expectation theorem*:

$$\boxed{\mathbb{E}\left(Y\right) = \mathbb{E}\left(\mathbb{E}\left(Y|X\right)\right)} \tag{7.22}$$

This equation is the fundamental relation of random (stochastic) processes. It relates the past/present (X) to the future Y, and thus is the basis for predicting the future.

7.3 Partitioning and Conditioning

We can understand and derive the *conditional expectation theorem* in a manner that is more transparent when applied to stochastic processes. A technique commonly used in calculation, based on this relation, is called *conditioning on* another variable, as the following discussion will explain.

Consider a partition of the sample space: $\{X_1, X_2, X_3, \ldots, X_n\}$, each event has a corresponding probability: $P(X_1), P(X_2), \ldots, P(X_n)$.

Consider another event Y, then according to the *partition rule* we have:

$$P(Y) = P(Y|X_1)P(X_1) + P(Y|X_2)P(X_2) + \cdots + P(Y|X_n)P(X_n) \quad . \tag{7.23}$$

Then it follows that:

$$\begin{aligned}
\mathbb{E}\left(Y\right) &= \sum_j y_j P(Y = y_j) && (7.24)\\
&= \sum_j y_j \left[P(Y = y_j|X_1)P(X_1) + \cdots + P(Y = y_j|X_n)P(X_n)\right] && (7.25)\\
&= \mathbb{E}\left(Y|X_1\right) P(X_1) + \mathbb{E}\left(Y|X_2\right) P(X_2) + \cdots + \mathbb{E}\left(Y|X_n\right) P(X_n) \quad . && (7.26)
\end{aligned}$$

Of course, we immediately recognise this as a special case of the *conditional expectation theorem*, which we repeat (without apology):

$$\mathbb{E}\left(Y\right) = \mathbb{E}\left(\mathbb{E}\left(Y|X\right)\right) \tag{7.27}$$

The following example will illustrate this idea. Suppose X is a Bernoulli variable, that is:

$$X = \{A, A^c\} \tag{7.28}$$

Then, according to the conditional expectation theorem, we have: $\mathbb{E}\left(Y\right) = \mathbb{E}\left(\mathbb{E}\left(Y|X\right)\right)$, this can be expanded in the form:

$$\boxed{\mathbb{E}\left(Y\right) = \mathbb{E}\left(Y|A\right) P(A) + \mathbb{E}\left(Y|A^c\right) P(A^c)} \tag{7.29}$$

The classic example of a Bernoulli trial is the tossing of a coin. Let us pose the following question. The probability of HEADS on a single toss of a coin is: $0 \leq p \leq 1$. What is the expected number of coin tosses before the first HEADS occurs ? We will present the solution to the problem with two approaches.

7.3.1 Solution 1: classic approach - summation

Let X denote the number of tosses required for the first HEADS; $X = \{1, 2, 3, \ldots, \infty\}$. If we have the probability mass for this variable, $f_X(x)$, it will be fairly simple to calculate its expectation by summation.

The probability that the first HEADS occurs on toss X, means that we require a succession of TAILS on toss 1, and toss 2, etc, up to toss $x-1$, followed by a HEADS on toss x. The probability of such an event is the geometric distribution:

$$f_X(x) = (1-p)^{x-1}p = q^{x-1}p \qquad x = 1, 2, 3, \ldots \qquad . \tag{7.30}$$

where, $q = 1 - p$, is the probability of TAILS on a single toss. Then it follows that:

$$\mathbb{E}(X) = \sum_{x=1}^{\infty} x\ q^{x-1}p \qquad . \tag{7.31}$$

This series can be summed using the usual device: $xq^{x-1} = (\partial/\partial q)q^x$. First. let's note that we can extend the range of the series to $x = 0$, since this adds nothing (the first term is zero):

$$\mathbb{E}(X) = \sum_{x=0}^{\infty} x\ .q^{x-1}p \tag{7.32}$$

Next apply the trick:

$$\mathbb{E}(X) = \sum_{x=0}^{\infty} p\frac{\partial}{\partial q}q^x = p\frac{\partial}{\partial q}\sum_{x=0}^{\infty} q^x \qquad . \tag{7.33}$$

The geometric series can be summed easily:

$$\mathbb{E}(X) = p\frac{\partial}{\partial q}\frac{1}{1-q} = p\frac{1}{(1-q)^2} = \frac{p}{p^2} = \frac{1}{p} \qquad . \tag{7.34}$$

This is the well-known result for the mean of a geometric distribution.

7.3.2 Solution 2: conditional expectation

Let event Y be the random outcome of the *first* flip of the coin. Then, "conditioning on Y", that is using (8.60), leads to the following expression:

$$\mathbb{E}(X) = \mathbb{E}(X|Y = \text{heads})\, P(Y = \text{heads}) + \mathbb{E}(X|Y = \text{tails})\, P(Y = \text{tails}) \tag{7.35}$$

That is:

$$\mathbb{E}(X) = \mathbb{E}(X|Y = H)\, p + \mathbb{E}(X|Y = T)\, q \qquad . \tag{7.36}$$

However, we can further simplify the right-hand-side as follows. If Y =heads, that is the first toss is given to be heads - then we can stop the process . The first heads has occured on the first toss, so that $X = 1$ for *certain*. X is not a random variable with this condition.

$$\mathbb{E}(X|Y = \text{heads}) = 1 \qquad . \tag{7.37}$$

For the other case, given that the first toss is tails then X must *certainly* be greater than 1. After the first game, we restart the whole process. So from this point the expected number of HEADS is exactly the same as it was before the first toss, since the conditions are identical, we have simply already tossed one coin. In mathematical terms, we write this as: .

$$\mathbb{E}(X|Y = \text{tails}) = 1 + \mathbb{E}(X) \qquad . \tag{7.38}$$

Thus:

$$\mathbb{E}(X) = 1.p. + (1 + \mathbb{E}(X))q \qquad . \tag{7.39}$$

This equation can be rearranged to find the value of $\mathbb{E}(X)$:

$$\mathbb{E}(X)(1-q) = p + q \quad , \tag{7.40}$$

leading to:

$$\mathbb{E}(X)\, p = 1 \qquad \Rightarrow \qquad \mathbb{E}(X) = \frac{1}{p} \qquad . \tag{7.41}$$

as before.

7.3.3 Conditional variance

In general we can define a *conditional variance* in a manner analogous to condition expectation:

$$\boxed{\text{var}\left(Y|X\right) \equiv \mathbb{E}\left(Y^2|X\right) - \left(\mathbb{E}\left(Y|X\right)\right)^2} \tag{7.42}$$

Let us pose an additional question for the coin tossing question. What is the variance of the number of coin tosses before the first HEADS occurs ?

We can *condition on* the first toss as before:

$$\mathbb{E}\left(X^2\right) = \mathbb{E}\left(X^2|Y = \text{heads}\right) P(Y = \text{heads}) + \mathbb{E}\left(X^2|Y = \text{tails}\right) P(Y = \text{tails}) \tag{7.43}$$

That is:

$$\mathbb{E}\left(X^2\right) = \mathbb{E}\left(X^2|Y = H\right) p + \mathbb{E}\left(X^2|Y = T\right) q \quad . \tag{7.44}$$

Now we have that:

$$\mathbb{E}\left(X^2|Y = H\right) = 1 \quad , \quad \mathbb{E}\left(X^2|Y = T\right) = \mathbb{E}\left((X+1)^2\right) \quad . \tag{7.45}$$

This gives us:

$$\mathbb{E}\left(X^2\right) = 1.p + (1 + 2\mathbb{E}\left(X\right) + \mathbb{E}\left(X^2\right))q = 1 + 2(q/p) + q\mathbb{E}\left(X^2\right) \quad , \tag{7.46}$$

with the result, after rearrangement,

$$\mathbb{E}\left(X^2\right) = \frac{1+q}{p^2} \quad . \tag{7.47}$$

It then follows that,

$$\text{var}(X) = \frac{(1+q)}{p^2} - \frac{1}{p^2} = \frac{q}{p^2} \quad , \tag{7.48}$$

giving us the variance in the number of tosses.

7.4 Conditional expectation for several variables

One can extend these ideas to three or more variables. In the following we will assume, for simplicity that we have 3 discrete random variables with a general probability mass function:

$$P(X = x, Y = y, Z = z) = P(x, y, z) \quad , \tag{7.49}$$

where we use some shorthand,$P(X = x_i, Y = y_j|Z = z_k) \equiv P(x, y|z)$, and so on.

This can be reduced to the the joint marginal mass:

$$P(x, y) \equiv \sum_z P(x, y, z) \quad ,$$

and similarly for $P(y, z)$ and $P(x, z)$. Then we have joint conditional probabilities, such as:

$$P(x, y|z) = \frac{P(x, y, z)}{P(z)} \quad , \tag{7.50}$$

and the conditional probability,

$$P(x|y, z) = \frac{P(x, y, z)}{P(y, z)} \quad . \tag{7.51}$$

Then we have the following useful lemmas:

- **Linearity** : $\mathbb{E}(aX + bY|Z) = a\mathbb{E}(X|Z) + b\mathbb{E}(Y|Z)$

  Proof:

$$\begin{aligned}\mathbb{E}(aX + bY|Z) &= \sum_{x,y}(ax + by)P(x, y|z) \\ &= a\sum_{x,y} xP(x, y|z) + b\sum_{x,y} yP(x, y|z) \\ &= a\sum_{x} xP(x|z) + b\sum_{y} yP(y|z) \\ &= a\mathbb{E}(X|Z) + b\mathbb{E}(Y|Z)\end{aligned}$$

(7.53)

- **Pull-through rule** :$\mathbb{E}(g(X)h(Y)|Y) = h(Y)\mathbb{E}(g(X)|Y)$

  Proof:

$$\begin{aligned}\mathbb{E}(g(X)h(Y)|Y = y) &= \sum_{x} g(x)h(y)P(x|y) \\ &= h(y)\sum_{x} g(x)P(x|y) \\ &= h(y)\mathbb{E}(g(X)|Y = y)\end{aligned}$$

- **Tower rule**: $\mathbb{E}(\mathbb{E}(Z|Y, X)|Y) = \mathbb{E}(\mathbb{E}(Z|Y)|Y, X) = \mathbb{E}(Z|Y)$

  This is a generalisation of the conditional expectation theorem.

  Proof:

$$\begin{aligned}\mathbb{E}(\mathbb{E}(Z|Y, X)|Y) &= \sum_{x}\left[\sum_{z} zP(z|y, x)\right]P(x|y) \\ &= \sum_{z} z\left[\sum_{x} P(z|y, x)P(x|y)\right] \\ &= \sum_{z} z\left[\sum_{x} P(x, y, z)/P(x, y) \times P(x, y)/P(y)\right] \\ &= \sum_{z} z\left[\sum_{x} P(x, y, z)/P(y)\right] \\ &= \sum_{z} zP(z|y) \\ &= \mathbb{E}(Z|Y)\end{aligned}$$

7.5 Predicting the future: regression

We already made allusion to the importance of conditional probability in terms of a *given* past and an uncertain future. One of the most important applications of stochastic processes is in prognostication (predicting the future). For example. suppose, we have a (time-ordered) sequence of data values

$$X_1, X_2, X_3, \ldots, X_n \quad ,$$

taken at times: $t_1, t_2, t_3, \ldots, t_n$. This corresponds to a set of *historical data*. We are uncertain whether these values are random or deterministic, whether they are correlated or not. Whatever the case, the problem of predicting the future is posed as follows. Given the historical record up to the present time t_n, we wish to predict the next number in the sequence, X_{n+1} at a future time t_{n+1}.

In a previous chapter, it was shown that the best predictor, in the sense of minimizing the expected least-squares error from a random variable in a set of independent (identical) trials, is the mean (expected) value. Similarly it can be shown that, given information,X, and a random variable Y, the best predictor (in the sense of minimizing the least-squares expected error) is the *conditional expectation* $\mathbb{E}(Y|X)$.

The essence of the problem is posed as follows. Given the value of a (random) variable $X = x$, which is the optimal function $s(X)$ that gives the best prediction of the value of another random variable, Y, given that Y may be correlated with the value of X ?

Consider the mean-square error between the values of Y, and the predictor $s(X)$:

$$g(s) \equiv \mathbb{E}\left((Y - s(X))^2 \,|X = x\right) \tag{7.54}$$

where, to be clear, this is an expectation is over the random variable, Y, which may depend on X.

As before, using the regression (conditional expectation) $\varphi_X(x) \equiv \mathbb{E}(Y|X = x)$, to play the role of the mean, we can write:

$$\begin{aligned} g(s) &\equiv \mathbb{E}\left((Y - \varphi_X(x) + \varphi_X(x) - s(x))^2 \,|X = x\right) && (7.55)\\ &= \mathbb{E}\left((Y - \varphi_X(x))^2 | X = x\right) && (7.56)\\ &\quad +2(\varphi_X(x) - s(x))\mathbb{E}(Y - \varphi_X(x)|X = x) + (\varphi_X(x) - s(x))^2 && (7.57) \end{aligned}$$

The first term on the right-hand-side is the conditional variance, which does not depend on our *guess*: $s(x)$. The middle term turns out to be zero, since, by definition, $\varphi_X(x) \equiv \mathbb{E}(Y|X = x)$. The only s-dependence is in the third term. This is minimal if, and only if: $s(x) - \varphi_X(x)$.

Thus, given information,X, and a random variable Y, the best predictor (in the sense of minimizing the least-squares expected error) is the *conditional* expectation, that is the *regression* of X on Y:

$$\boxed{\text{Best predictor} = \mathbb{E}(Y|X) = \varphi_X(x)} \tag{7.58}$$

It is no exaggeration to say that *regression* is one of the fundamental tools in a statisticians box of tricks. Some would say it is the main technique of statistical modelling. The simple *linear model* of statistics is an example of regression, in this case the least-squares linear regression. This can be extended to *generalised linear models* of regression. Regression is perhaps the best-known black-box statistical tool used across the sciences. However, outside of the mathematically-inclined practitoners, very few users know what it is (or care for that matter), and even fewer know how to calculate it!

7.5.1 Time series and constructing models

Returning to the time series question, the best predictor of X_{n+1} (in the sense of minimizing the least-squares expected error), given the historical record: $X_1, \ldots, X_n$, is:

$$\hat{X}_{n+1} = \mathbb{E}(X_{n+1}|X_1, X_2, X_3, \ldots, X_n) \qquad . \tag{7.59}$$

Now, here's the tricky part. This is wonderful in theory but, in most cases, *one simply doesn't know* the conditional probability mass function: $f_{X_{n+1}|X_1,X_2,X_3,\ldots,X_n}(x_{n+1}; x_1, x_2, x_3, \ldots, x_n)$. As is generally the case one does not know even the random process behind the data. In that case one tries to *construct a model*. In simple terms, attempt to guess a functional form for the probability density. *Time-series analysis* is an attempt to extract the deterministic (correlated) connection between the past and the future, and at the same time to use the historical record to determine the nature and degree of randomness (the probability mass function if you like).

The next chapter is devoted to a time series. A classic problem called the *simple random walk*. In this application, we will see some of the theory developed thus far, put into practice.