

## 5. Conditional Expected Value

As usual, our starting point is a [random experiment](#) with [probability measure](#)  $\mathbb{P}$  on a [sample space](#)  $\Omega$ . Suppose that  $X$  is a [random variable](#) taking values in a set  $S$  and that  $Y$  is a random variable taking values in  $T \subseteq \mathbb{R}$ . In this section, we will study the conditional expected value of  $Y$  given  $X$ , a concept of fundamental importance in probability. As we will see, the expected value of  $Y$  given  $X$  is the function of  $X$  that best approximates  $Y$  in the mean square sense. Note that in general,  $X$  will be vector-valued. In this section, we will assume that all real-valued random variables occurring in expected values have finite second moment.

### Basic Theory

#### The Elementary Definition

Note that we can think of  $(X, Y)$  as a random variable that takes values in a subset of  $S \times T$ . Suppose first that the that  $(X, Y)$  has a (joint) [continuous distribution](#) with probability density function  $f$ . Recall that the [\(marginal\) probability density function](#)  $g$  of  $X$  is given by

$$g(x) = \int_T f(x, y) dy, \quad x \in S$$

and that the [conditional probability density function](#) of  $Y$  given  $X = x$  is given by

$$h(y|x) = \frac{f(x, y)}{g(x)}, \quad x \in S, y \in T$$

Thus, the [conditional expected value](#) of  $Y$  given  $X = x$  is simply the mean computed relative to the conditional distribution:

$$\mathbb{E}(Y|X = x) = \int_T y h(y|x) dy, \quad x \in S$$

Of course, the conditional mean of  $Y$  depends on the given value  $x$  of  $X$ . Temporarily, let  $v$  denote the function from  $S$  into  $\mathbb{R}$  defined by

$$v(x) = \mathbb{E}(Y|X = x), \quad x \in S$$

The function  $v$  is sometimes referred to as the [regression function](#) of  $Y$  based on  $X$ . The random variable  $v(X)$  is called the [conditional expected value](#) of  $Y$  given  $X$  and is denoted  $\mathbb{E}(Y|X)$ . The results and definitions above would be exactly the same if  $(X, Y)$  has a joint [discrete distribution](#), except that sums would replace the integrals. Intuitively, we treat  $X$  as known, and therefore not random, and we then average  $Y$  with respect to the probability distribution that remains.

## The General Definition

The random variable  $\mathbb{E}(Y|X)$  satisfies a critical property that characterizes it among all functions of  $X$ .

- ❖ 1. Suppose that  $r$  is a function from  $S$  into  $\mathbb{R}$ . Use the [change of variables theorem](#) for expected value to show that

$$\mathbb{E}(r(X) \mathbb{E}(Y|X)) = \mathbb{E}(r(X) Y)$$

In fact, the result in [Exercise 1](#) can be used as a definition of conditional expected value, regardless of the type of the distribution of  $(X, Y)$ . Thus, generally we *define*  $\mathbb{E}(Y|X)$  to be the random variable that satisfies the condition in [Exercise 1](#) and is of the form  $\mathbb{E}(Y|X) = v(X)$  for some function  $v$  from  $S$  into  $\mathbb{R}$ . Then we define  $\mathbb{E}(Y|X = x)$  to be  $v(x)$  for  $x \in S$ . (More technically,  $\mathbb{E}(Y|X)$  is required to be [measurable](#) with respect to  $X$ .)

## Properties

Our first consequence of [Exercise 1](#) is a formula for computing the expected value of  $Y$ .

- ❖ 2. By taking  $r$  to be the constant function 1 in [Exercise 1](#), show that  $\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y)$

Aside from the theoretical interest, the result in [Exercise 2](#) is often a good way to compute  $\mathbb{E}(Y)$  when we know the conditional distribution of  $Y$  given  $X$ . We say that we are computing the expected value of  $Y$  by [conditioning](#) on  $X$ .

- ❖ 3. Show that, in light of [Exercise 2](#), the condition in [Exercise 1](#) can be restated as follows: For any function  $r : S \rightarrow \mathbb{R}$ , the random variables  $Y - \mathbb{E}(Y|X)$  and  $r(X)$  are uncorrelated.

The next exercise show that the condition in [Exercise 1](#) characterizes  $\mathbb{E}(Y|X)$

- ❖ 4. Suppose that  $u(X)$  and  $v(X)$  satisfy the condition in [Exercise 1](#) and hence also the results in [Exercise 2](#) and [Exercise 3](#). Show that  $u(X)$  and  $v(X)$  are [equivalent](#):

- First show that  $\text{var}(u(X) - v(X)) = 0$
- Then show that  $\mathbb{P}(u(X) = v(X)) = 1$ .

- ❖ 5. Suppose that  $s : S \rightarrow \mathbb{R}$ . Use the characterization in [Exercise 1](#) to show that

$$\mathbb{E}(s(X) Y|X) = s(X) \mathbb{E}(Y|X)$$

This result makes intuitive sense: if we know  $X$ , then we know any (deterministic) function of  $X$ . Any such function acts like a constant in terms of the conditional expected value with respect to  $X$ . The following rule generalizes this result and is sometimes referred to as the [substitution rule](#) for conditional expected value.

- ❖ 6. Suppose that  $s : S \times T \rightarrow \mathbb{R}$ . Show that

$$\mathbb{E}(s(X, Y)|X = x) = \mathbb{E}(s(x, Y)|X = x)$$

In particular, it follows from [Exercise 5](#) or [Exercise 6](#) that  $\mathbb{E}(s(X)|X) = s(X)$ . At the opposite extreme, we have the result in the next exercise. If  $X$  and  $Y$  are independent, then knowledge of  $X$  gives no information about  $Y$  and so the conditional expected value with respect to  $X$  is the same as the ordinary (unconditional) expected value of  $Y$ .

7. Suppose that  $X$  and  $Y$  are independent. Use the characterization in [Exercise 1](#) to show that

$$\mathbb{E}(Y|X) = \mathbb{E}(Y)$$

Use the general definition to establish the properties in the following exercises, where  $Y$  and  $Z$  are real-valued random variables and  $c$  is a constant. Note that these are analogies of basic properties of ordinary expected value. Every type of expected value must satisfy two critical properties: linearity and monotonicity.

8. Show that  $\mathbb{E}(Y + Z|X) = \mathbb{E}(Y|X) + \mathbb{E}(Z|X)$ .

9. Show that  $\mathbb{E}(cY|X) = c\mathbb{E}(Y|X)$ .

10. Show that if  $Y \geq 0$  with probability 1, then  $\mathbb{E}(Y|X) \geq 0$  with probability 1.

11. Show that if  $Y \leq Z$  with probability 1, then  $\mathbb{E}(Y|X) \leq \mathbb{E}(Z|X)$  with probability 1.

12. Show that  $|\mathbb{E}(Y|X)| \leq \mathbb{E}(|Y||X)$  with probability 1.

Suppose now that  $Z$  is real-valued and that  $X$  and  $Y$  are random variables (all defined on the same probability space, of course). The following exercise gives a consistency condition of sorts. Iterated conditional expected values reduce to a single conditional expected value with respect to the minimum amount of information:

13. Show that

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

## Conditional Probability

The conditional probability of an event  $A$ , given random variable  $X$ , is a special case of the conditional expected value. As usual, let  $\mathbf{1}(A)$  denote the indicator random variable of  $A$ . We define

$$\mathbb{P}(A|X) = \mathbb{E}(\mathbf{1}(A)|X)$$

The properties above for conditional expected value, of course, have special cases for conditional probability.

14. Show that  $\mathbb{P}(A) = \mathbb{E}(\mathbb{P}(A|X))$ .

Again, the result in the previous exercise is often a good way to compute  $\mathbb{P}(A)$  when we know the conditional probability of  $A$  given  $X$ . We say that we are computing the probability of  $A$  by **conditioning** on  $X$ . This is a very compact and elegant version of the conditioning result given first in the section on [Conditional Probability](#) in the chapter on [Probability Spaces](#) and later in the section on [Discrete Distributions](#) in the Chapter on [Distributions](#).

## The Best Predictor

The next two exercises show that, of all functions of  $X$ ,  $\mathbb{E}(Y|X)$  is the best predictor of  $Y$ , in the sense of minimizing the mean square error. This is fundamentally important in statistical problems where the **predictor vector**  $X$  can be observed but not the **response variable**  $Y$ .

15. Let  $v(X) = \mathbb{E}(Y|X)$  and let  $u : S \rightarrow \mathbb{R}$ . By adding and subtracting  $v(X)$ , expanding, and using the result of Exercise 3, show that

$$\mathbb{E}((Y - u(X))^2) = \mathbb{E}((Y - v(X))^2) + \mathbb{E}((v(X) - u(X))^2)$$

16. Use the result of the last exercise to show that if  $u : S \rightarrow \mathbb{R}$ , then

$$\mathbb{E}((\mathbb{E}(Y|X) - Y)^2) \leq \mathbb{E}((u(X) - Y)^2)$$

and equality holds if and only if  $u(X) = \mathbb{E}(Y|X)$  with probability 1.

Suppose now that  $X$  is real-valued. In the section on [covariance and correlation](#), we found that the best *linear* predictor of  $Y$  based on  $X$  is

$$L(Y|X) = \mathbb{E}(Y) + \frac{\text{cov}(X, Y)}{\text{var}(X)}(X - \mathbb{E}(X))$$

On the other hand,  $\mathbb{E}(Y|X)$  is the best predictor of  $Y$  among *all* functions of  $X$ . It follows that if  $\mathbb{E}(Y|X)$  happens to be a linear function of  $X$  then it must be the case that  $\mathbb{E}(Y|X) = L(Y|X)$

17. Show that  $\text{cov}(X, \mathbb{E}(Y|X)) = \text{cov}(X, Y)$

18. Show directly that if  $\mathbb{E}(Y|X) = aX + b$  then

a.  $a = \frac{\text{cov}(X, Y)}{\text{var}(X)}$

b.  $b = \mathbb{E}(Y) - \frac{\text{cov}(X, Y)}{\text{var}(X)}\mathbb{E}(X)$

## Conditional Variance

The **conditional variance** of  $Y$  given  $X$  is naturally defined as follows:

$$\text{var}(Y|X) = \mathbb{E}((Y - \mathbb{E}(Y|X))^2 | X)$$

19. Show that  $\text{var}(Y|X) = \mathbb{E}(Y^2 | X) - \mathbb{E}(Y|X)^2$ .

20. Show that  $\text{var}(Y) = \mathbb{E}(\text{var}(Y|X)) + \text{var}(\mathbb{E}(Y|X))$ .

Again, the result in the previous exercise is often a good way to compute  $\text{var}(Y)$  when we know the

conditional distribution of  $Y$  given  $X$ . We say that we are computing the variance of  $Y$  by **conditioning on  $X$** .

Let us return to the study of predictors of the real-valued random variable  $Y$ , and compare the three predictors we have studied in terms of mean square error.

1. First, the best **constant predictor** of  $Y$  is  $\mu = \mathbb{E}(Y)$  with mean square error  $\text{var}(Y) = \mathbb{E}((Y - \mu)^2)$
2. Next, if  $X$  is another real-valued random variable, then as we showed in the section on **covariance and correlation**, the best **linear predictor** of  $Y$  based on  $X$  is

$$L(Y|X) = \mathbb{E}(Y) + \frac{\text{cov}(X, Y)}{\text{var}(X)} (X - \mathbb{E}(X))$$

with mean square error  $\mathbb{E}((Y - L(Y|X))^2) = \text{var}(Y) (1 - \text{cor}(X, Y)^2)$ .

3. Finally, if  $X$  is a general random variable, then as we have shown in this section, the best **overall predictor** of  $Y$  based on  $X$  is  $\mathbb{E}(Y|X)$  with mean square error  $\mathbb{E}(\text{var}(Y|X)) = \text{var}(Y) - \text{var}(\mathbb{E}(Y|X))$ .

## Examples and Applications

21. Suppose that  $(X, Y)$  has probability density function  $f(x, y) = x + y$ ,  $0 \leq x \leq 1$ ,  $0 \leq y \leq 1$ .

- a. Find  $L(Y|X)$ .
- b. Find  $\mathbb{E}(Y|X)$ .
- c. Graph  $L(Y|X = x)$  and  $\mathbb{E}(Y|X = x)$  as functions of  $x$ , on the same axes.
- d. Find  $\text{var}(Y)$ .
- e. Find  $\text{var}(Y) (1 - \text{cor}(X, Y)^2)$ .
- f. Find  $\text{var}(Y) - \text{var}(\mathbb{E}(Y|X))$ .



22. Suppose that  $(X, Y)$  has probability density function  $f(x, y) = 2(x + y)$ ,  $0 \leq x \leq y \leq 1$ .

- a. Find  $L(Y|X)$ .
- b. Find  $\mathbb{E}(Y|X)$ .
- c. Graph  $L(Y|X = x)$  and  $\mathbb{E}(Y|X = x)$  as functions of  $x$ , on the same axes.
- d. Find  $\text{var}(Y)$ .
- e. Find  $\text{var}(Y) (1 - \text{cor}(X, Y)^2)$ .
- f. Find  $\text{var}(Y) - \text{var}(\mathbb{E}(Y|X))$ .



23. Suppose that  $(X, Y)$  has probability density function  $f(x, y) = 6x^2y$ ,  $0 \leq x \leq 1$ ,  $0 \leq y \leq 1$ .

- a. Find  $L(Y|X)$ .
- b. Find  $\mathbb{E}(Y|X)$ .
- c. Graph  $L(Y|X = x)$  and  $\mathbb{E}(Y|X = x)$  as functions of  $x$ , on the same axes.

- d. Find  $\text{var}(Y)$ .
- e. Find  $\text{var}(Y) (1 - \text{cor}(X, Y)^2)$ .
- f. Find  $\text{var}(Y) - \text{var}(\mathbb{E}(Y|X))$ .



24. Suppose that  $(X, Y)$  has probability density function  $f(x, y) = 15x^2y$ ,  $0 \leq x \leq y \leq 1$ .

- a. Find  $L(Y|X)$ .
- b. Find  $\mathbb{E}(Y|X)$ .
- c. Graph  $L(Y|X = x)$  and  $\mathbb{E}(Y|X = x)$  as functions of  $x$ , on the same axes.
- d. Find  $\text{var}(Y)$ .
- e. Find  $\text{var}(Y) (1 - \text{cor}(X, Y)^2)$ .
- f. Find  $\text{var}(Y) - \text{var}(\mathbb{E}(Y|X))$ .



25. Suppose that  $X, Y$ , and  $Z$  are real-valued random variables with  $\mathbb{E}(Y|X) = X^3$  and  $\mathbb{E}(Z|X) = \frac{1}{1+X^2}$ . Find  $\mathbb{E}(Y e^X - Z \sin(X)|X)$ .



### Uniform Distributions

26. Suppose that  $(X, Y)$  is uniformly distributed on a rectangular region  $S = [a, b] \times [c, d] \subseteq \mathbb{R}^2$ . Find  $\mathbb{E}(Y|X)$ .



27. In the **bivariate uniform experiment**, select the *square* in the list box. Run the simulation 2000 times, updating every 10 runs. Note the relationship between the cloud of points and the graph of the regression function.

28. Suppose that  $(X, Y)$  is uniformly distributed on a triangular region  $T = \{(x, y) \in \mathbb{R}^2 : -a \leq x \leq y \leq a\}$ , where  $a > 0$  is a parameter. Find  $\mathbb{E}(Y|X)$ .



29. In the **bivariate uniform experiment**, select the *triangle* in the list box. Run the simulation 2000 times, updating every 10 runs. Note the relationship between the cloud of points and the graph of the regression function.

30. Suppose that  $X$  is uniformly distributed on  $[0, 1]$ , and that given  $X$ ,  $Y$  is uniformly distributed on  $[0, X]$ . Find each of the following:

- a.  $\mathbb{E}(Y|X)$
- b.  $\mathbb{E}(Y)$

- c.  $\text{var}(Y|X)$
- d.  $\text{var}(Y)$



## Coins and Dice

31. A pair of fair dice are thrown, and the scores  $(X_1, X_2)$  recorded. Let  $Y = X_1 + X_2$  denote the sum of the scores and  $U = \min \{X_1, X_2\}$  the minimum score. Find each of the following:

- a.  $\mathbb{E}(Y|X_1)$
- b.  $\mathbb{E}(U|X_1)$
- c.  $\mathbb{E}(Y|U)$
- d.  $\mathbb{E}(X_2|X_1)$



32. A box contains 10 coins, labeled 0 to 9. The probability of heads for coin  $i$  is  $\frac{i}{9}$ . A coin is chosen at random from the box and tossed. Find the probability of heads. This problem is an example of [Laplace's rule of succession](#),



## Random Sums of Random Variables

Suppose that  $X = (X_1, X_2, \dots)$  is a sequence of independent and identically distributed real-valued random variables. We will denote the common mean, variance, and moment generating function, respectively, by:  $\mu = \mathbb{E}(X_i)$ ,  $\sigma^2 = \text{var}(X_i)$ , and  $G(t) = \mathbb{E}(e^{tX_i})$ . Let

$$Y_n = \sum_{i=1}^n X_i, \quad n \in \mathbb{N}$$

so that  $Y = (Y_0, Y_1, \dots)$  is the partial sum process associated with  $X$ . Suppose now that  $N$  is a random variable taking values in  $\mathbb{N}$ , independent of  $X$ . Then

$$Y_N = \sum_{i=1}^N X_i$$

is a random sum of random variables; the terms in the sum are random, and the number of terms is random. This type of variable occurs in many different contexts. For example,  $N$  might represent the number of customers who enter a store in a given period of time, and  $X_i$  the amount spent by the customer  $i$ .

33. Show that

- $\mathbb{E}(Y_N|N) = N\mu$
- $\mathbb{E}(Y_N) = \mathbb{E}(N)\mu$

**Wald's equation**, named for **Abraham Wald**, is a generalization of the result in the previous exercise to the case where  $N$  is not necessarily independent of  $\mathbf{X}$ , but rather is a **stopping time** for  $\mathbf{X}$ . **Wald's equation** is discussed in the section on **Partial Sums** in the chapter on **Random Samples**.

34. Show that

- $\text{var}(Y_N|N) = N\sigma^2$
- $\text{var}(Y_N) = \mathbb{E}(N)\sigma^2 + \text{var}(N)\mu^2$

35. Let  $H$  denote the probability generating function of  $N$ . Show that the moment generating function of  $Y_N$  is  $H \circ G$ .

- $\mathbb{E}(e^{tY_N}|N) = G(t)^N$
- $\mathbb{E}(e^{tY_N}) = H(G(t))$

36. In the die-coin experiment, a fair die is rolled and then a fair coin is tossed the number of times showing on the die. Let  $N$  denote the die score and  $Y$  the number of heads.

- Find the conditional distribution of  $Y$  given  $N$ .
- Find  $\mathbb{E}(Y|N)$ .
- Find  $\text{var}(Y|N)$ .
- Find  $\mathbb{E}(Y)$ .
- Find  $\text{var}(Y)$ .



37. Run the **die-coin experiment** 1000 times, updating every 10 runs. Note the apparent convergence of the empirical mean and standard deviation to the distribution mean and standard deviation.

38. The number of customers entering a store in a given hour is a random variable with mean 20 and standard deviation 3. Each customer, independently of the others, spends a random amount of money with mean \$50 and standard deviation \$5. Find the mean and standard deviation of the amount of money spent during the hour.



39. A coin has a random probability of heads  $V$  and is tossed a random number of times  $N$ . Suppose that  $V$  is uniformly distributed on  $[0, 1]$ ;  $N$  has the Poisson distribution with parameter  $a > 0$ ; and  $V$  and  $N$  are independent. Let  $Y$  denote the number of heads. Compute the following:

- $\mathbb{E}(Y|N, V)$
- $\mathbb{E}(Y|N)$
- $\mathbb{E}(Y|V)$
- $\mathbb{E}(Y)$
- $\text{var}(Y|N, V)$
- $\text{var}(Y)$



## Mixtures of Distributions

Suppose that  $\mathbf{X} = (X_1, X_2, \dots)$  is a sequence of real-valued random variables. Let  $\mu_i = \mathbb{E}(X_i)$ ,  $\sigma_i^2 = \text{var}(X_i)$ , and  $M_i(t) = \mathbb{E}(e^{tX_i})$  for  $i \in \mathbb{N}_+$ . Suppose also that  $N$  is a random variable taking values in  $\mathbb{N}_+$ , independent of  $\mathbf{X}$ . Denote the probability density function of  $N$  by  $p_i = \mathbb{P}(N = i)$  for  $i \in \mathbb{N}_+$ . The distribution of the random variable  $X_N$  is a **mixture** of the distributions of  $\mathbf{X} = (X_1, X_2, \dots)$ , with the distribution of  $N$  as the mixing distribution.

40. Show that  $\mathbb{E}(X_N|N) = \mu_N$

41. Show that  $\mathbb{E}(X_N) = \sum_{i=1}^{\infty} p_i \mu_i$ .

42. Show that  $\text{var}(X_N) = \sum_{i=1}^{\infty} p_i (\sigma_i^2 + \mu_i^2) - \sum_{i=1}^{\infty} p_i \mu_i^2$ .

43. Show that  $\mathbb{E}(e^{tX_N}) = \sum_{i=1}^{\infty} p_i M_i(t)$

44. In the coin-die experiment, a biased coin is tossed with probability of heads  $\frac{1}{3}$ . If the coin lands tails, a fair die is rolled; if the coin lands heads, an ace-six flat die is rolled (faces 1 and 6 have probability  $\frac{1}{4}$  each, and faces 2, 3, 4, 5 have probability  $\frac{1}{8}$  each). Find the mean and standard deviation of the die score.



45. Run the **coin-die experiment** 1000 times, updating every 10 runs. Note the apparent convergence of the empirical mean and standard deviation to the distribution mean and standard deviation.

## Vector Space Concepts

Conditional expectation can be interpreted in terms vector of space concepts. This connection can help illustrate many of the properties of conditional expectation from a different point of view.

Recall that the **vector space**  $\mathcal{V}_2$  consists of all real-valued random variables defined on a fixed sample space  $\Omega$  (that is, relative to the same random experiment) that have finite second moment. Recall that two random variables are **equivalent** if they are equal with probability 1. We consider two such random variables as the same vector, so that technically, our vector space consists of **equivalence classes** under this **equivalence**

**relation**. The **addition operator** corresponds to the usual addition of two real-valued random variables, and the operation of **scalar multiplication** corresponds to the usual multiplication of a real-valued random variable by a real (non-random) number.

Recall also that  $\mathcal{V}_2$  is an **inner product space**, with inner product given by

$$\langle U, V \rangle = \mathbb{E}(UV)$$

## Projections

Suppose now that  $X$  a general random variable defined on the sample space  $S$ , and that  $Y$  is a real-valued random variable in  $\mathcal{V}_2$ .

46. Show that the set below is a subspace of  $\mathcal{V}_2$ :

$$\mathcal{U} = \{U \in \mathcal{V}_2 : U = u(X) \text{ for some } u : S \rightarrow \mathbb{R}\}$$

47. Reconsider [Exercise 3](#) to show that  $\mathbb{E}(Y|X)$  is the **projection** of  $Y$  on to the subspace  $\mathcal{U}$ .

Suppose now that  $X \in \mathcal{V}_2$ . Recall that the set

$$\mathcal{W} = \{W \in \mathcal{V}_2 : W = aX + b \text{ for some } (a \in \mathbb{R}) \text{ and } (b \in \mathbb{R})\}$$

is also a subspace of  $\mathcal{V}_2$ , and in fact is clearly also a subspace of  $\mathcal{U}$ . We showed that the  $L(Y|X)$  is the projection of  $Y$  onto  $\mathcal{W}$ .