HarvardX: STAT110x
Introduction to Probability

Help sandipan_dey

Course > Unit 6: Joint Distributions and Conditional Expectation > 6.1 Reading > 6.5 Conditional expectation

6.5 Conditional expectation

Unit 6: Joint Distributions and Conditional Expectation

Adapted from Blitzstein-Hwang Chapters 7 and 9.

As you might guess, conditional expectation is expectation, with conditional probability in place of probability. This is an essential concept, for reasons analogous to why we need conditional probability:

- Conditional expectation is a powerful tool for calculating expectations. Using strategies such as conditioning on what we wish we knew and first-step analysis, we can often decompose complicated expectation problems into simpler pieces.
- Conditional expectation is a relevant quantity in its own right, allowing us to predict or estimate unknowns based on whatever evidence is currently available.

There are two different but closely linked notions of conditional expectation:

- *Conditional expectation* $E(Y|A)$ *given an* $\{\text{event}\}$: let Y be an r.v., and A be an event. If we learn that A occurred, our updated expectation for Y, $E(Y|A)$, is computed analogously to $E(Y)$, except using conditional probabilities given A.
- *Conditional expectation* $E(Y|X)$ *given a* $\{\text{random variable}\}$: a more subtle question is how to define $E(Y|X)$, where X and Y are both r.v.s. Intuitively, $E(Y|X)$ is the r.v. that best predicts Y using only the information available from X.

Recall that the expectation $E(Y)$ of a discrete r.v. Y is a weighted average of its possible values, where the weights are the PMF values $P(Y=y)$. After learning that an event A occurred, we want to use weights that have been updated to reflect this new information. The definition of $E(Y|A)$ simply replaces the probability $P(Y=y)$ with the conditional probability $P(Y=y|A)$.

Similarly, if Y is continuous, $E(Y)$ is still a weighted average of the possible values of Y, with an integral in place of a sum and the PDF value $f(y)$ in place of a PMF value. If we learn that A occurred, we update the expectation for Y by replacing $f(y)$ with the conditional PDF $f(y|A)$.

Definition 6.5.1 (Conditional expectation given an event).

Let A be an event with positive probability. If Y is a discrete r.v., then the *conditional expectation of* Y *given* A is

$$E(Y|A) = \sum_y yP(Y=y|A),$$

where the sum is over the support of Y. If Y is a continuous r.v. with PDF f, then

$$E(Y|A) = \int_{-\infty}^{\infty} yf(y|A)dy,$$

where the conditional PDF $f(y|A)$ is defined as the derivative of the conditional CDF $F(y|A) = P(Y \leq y|A)$, and can also be computed by a hybrid version of Bayes' rule:

$$f(y|A) = \frac{P(A|Y = y)f(y)}{P(A)}.$$

☣ **Warning 6.5.2.**

Confusing conditional expectation and unconditional expectation is a dangerous mistake. More generally, not keeping careful track of what you *should be* conditioning on and what you *are* conditioning on is a recipe for disaster.

Example 6.5.3 (Life expectancy).

Fred is 30 years old, and he hears that the average life expectancy in his country is 80 years. Should he conclude that, on average, he has 50 years of life left? No, there is a crucial piece of information that he must condition on: the fact that he has lived to age 30 already. Letting T be Fred's lifespan, we have the cheerful news that

$$E(T) < E(T|T \geq 30).$$

The left-hand side is Fred's life expectancy at birth (it implicitly conditions on the fact that he is born), and the right-hand side is Fred's life expectancy given that he reaches age 30.

The law of total probability allows us to get unconditional probabilities by slicing up the sample space and computing conditional probabilities in each slice. The same idea works for computing unconditional expectations.

Theorem 6.5.4 (Law of total expectation).

Let $A_1, \ldots, A_n$ be a partition of a sample space, with $P(A_i) > 0$ for all i, and let Y be a random variable on this sample space. Then

$$E(Y) = \sum_{i=1}^{n} E(Y|A_i)P(A_i).$$

In fact, since all probabilities are expectations by the fundamental bridge, the law of total probability is a special case of the law of total expectation. To see this, let $Y = I_B$ for an event B; then the above theorem says

$$P(B) = E(I_B) = \sum_{i=1}^{n} E(I_B|A_i)P(A_i) = \sum_{i=1}^{n} P(B|A_i)P(A_i),$$

which is exactly LOTP. The law of total expectation is, in turn, a special case of a major result called *Adam's law* (Theorem 6.7.1), so we will not prove it yet. There are many interesting examples of using wishful thinking to break up an unconditional expectation into conditional expectations. We begin with two cautionary tales about the importance of conditioning carefully and not destroying information without justification.

Example 6.5.5 (Two-envelope paradox).

A stranger presents you with two identical-looking, sealed envelopes, each of which contains a check for some positive amount of money. You are informed that one of the envelopes contains exactly twice as much money as the other. You can choose either envelope. Which do you prefer: the one on the left or the one on the right? (Assume that the expected amount of money in each envelope is finite---certainly a good assumption in the real world!)

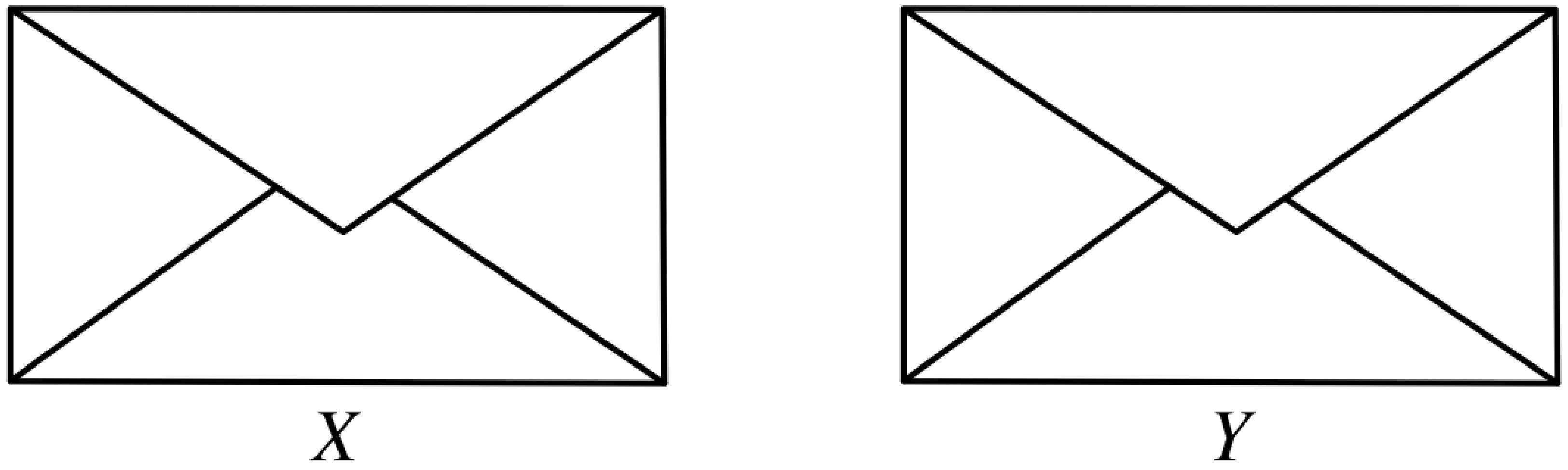

Figure 6.5.6: Two envelopes, where one contains twice as much money as the other. Either $Y = 2X$ or $Y = X/2$, with equal probabilities. Which would you prefer?

View Larger Image
Image Description

Solution

Let X and Y be the amounts in the left and right envelopes, respectively. By symmetry, there is no reason to prefer one envelope over the other (we are assuming there is no prior information that the stranger is left-handed and left-handed people prefer putting more money on the left). Concluding by symmetry that $E(X) = E(Y)$, it seems that you should not care which envelope you get.

But as you daydream about what's inside the envelopes, another argument occurs to you: suppose that the left envelope has \$100. Then the right envelope either has \$50 or \$200. The average of \$50 and \$200 is \$125, so it seems then that the right envelope is better. But there was nothing special about \$100 here; for any value x for the left envelope, the average of $2x$ and $x/2$ is greater than x, suggesting that the right envelope is better. This is bizarre though, since not only does it contradict the symmetry argument, but also the same reasoning could be applied starting with the right envelope, leading to switching back and forth forever!

Let us try to formalize this argument to see what's going on. We have $Y = 2X$ or $Y = X/2$, with equal probabilities. By Theorem 6.5.4,

$$E(Y) = E(Y|Y = 2X) \cdot \frac{1}{2} + E\big(Y|Y = X/2\big) \cdot \frac{1}{2}.$$

One might then think that this is

$$E(2X) \cdot \frac{1}{2} + E\big(X/2\big) \cdot \frac{1}{2} = \frac{5}{4}E(X),$$

suggesting a 25 gain from switching from the left to the right envelope. But there is a blunder in that calculation: $E(Y|Y = 2X) = E(2X|Y = 2X)$, but there is no justification for dropping the $Y = 2X$ condition after plugging in $2X$ for Y.

To put it another way, let I be the indicator of the event $Y = 2X$, so that $E(Y|Y = 2X) = E(2X|I = 1)$. If we know that X is independent of I, then we can drop the condition $I = 1$. But in fact we have just *proven* that X and I can't be independent: if they were, we'd have a paradox! Surprisingly, *observing X gives information about whether X is the bigger value or the smaller value*!

The next example vividly illustrates the importance of conditioning on *all* the information. The phenomenon revealed here arises in many real-life decisions about what to buy and what investments to make.

Example 6.5.7 (Mystery prize).

You are approached by another stranger, who gives you an opportunity to bid on a mystery box containing a mystery prize! The value of the prize is completely unknown, except that it is worth at least nothing, and at most a million dollars. So the true value V of the prize is considered to be Uniform on [0,1] (measured in millions of dollars).

You can choose to bid any amount b (in millions of dollars). You have the chance to get the prize for considerably less than it is worth, but you could also lose money if you bid too much. Specifically, if $b < 2V/3$, then the bid is rejected and nothing is gained or lost. If $b \geq 2V/3$, then the bid is accepted and your net payoff is $V - b$ (since you pay b to get a prize worth V). What is your optimal bid b, to maximize the expected payoff?

Solution

Your bid $b \geq 0$ must be a predetermined constant (not based on V, since V is unknown!). To find the expected payoff W, condition on whether the bid is accepted. The payoff is $V - b$ if the bid is accepted and 0 if the bid is rejected. So

$$\begin{aligned} E(W) &= E(W|b \geq 2V/3)P(b \geq 2V/3) + E(W|b < 2V/3)P(b < 2V/3) \\ &= E(V - b|b \geq 2V/3)P(b \geq 2V/3) + 0 \\ &= \left(E(V|V \leq 3b/2) - b\right) P(V \leq 3b/2). \end{aligned}$$

For $b \geq 2/3$, the event $V \leq 3b/2$ has probability 1, so the right-hand side is $1/2 - b < 0$, i.e., you lose money on average. Now assume $b < 2/3$. Then $V \leq 3b/2$ has probability $3b/2$. Given that $V \leq 3b/2$, the conditional distribution of V is Uniform on $[0, 3b/2]$. Therefore,

$$E(W) = (E(V|V \leq 3b/2) - b)\, P(V \leq 3b/2) = (3b/4 - b)\,(3b/2) = -3b^2/8.$$

The above expression is negative except at $b = 0$, so the optimal bid is 0: you shouldn't play this game!

Example 6.5.8 (Time until *HH* vs. *HT*).

You toss a fair coin repeatedly. What is the expected number of tosses until the pattern *HT* appears for the first time? What about the expected number of tosses until *HH* appears for the first time?

Solution

Let W_{HT} be the number of tosses until *HT* appears. As we can see from Figure 6.5.9, W_{HT} is the waiting time for the first Heads, which we'll call W_1, plus the additional waiting time for the first Tails after the first Heads, which we'll call W_2. By the story of the First Success distribution, W_1 and W_2 are i.i.d. $\mathrm{FS}(1/2)$, so $E(W_1) = E(W_2) = 2$ and $E(W_{HT}) = 4$.

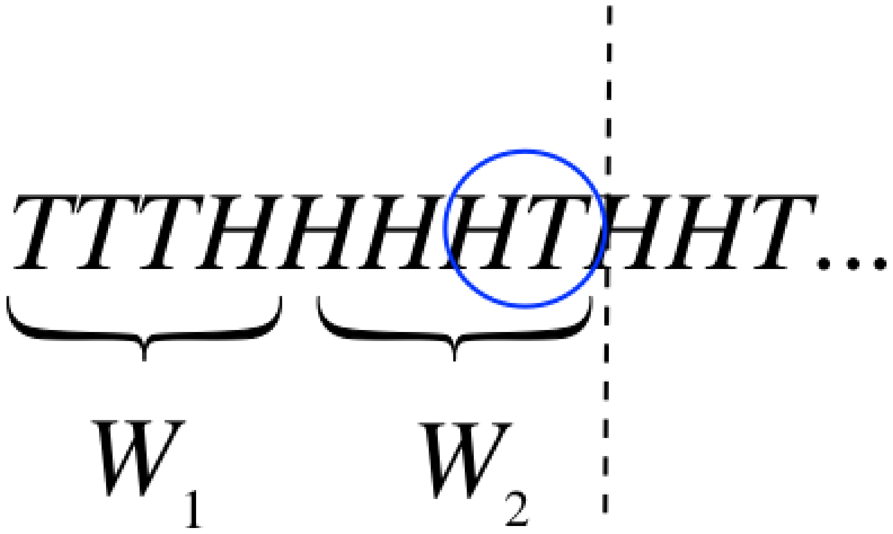

Figure 6.5.9: Waiting time for HT is the waiting time for the first Heads, W_1, plus the additional waiting time for the next Tails, W_2. Durable partial progress is possible!

View Larger Image
Image Description

Finding the expected waiting time for *HH*, $E(W_{HH})$, is more complicated. We can't apply the same logic as for $E(W_{HT})$: as shown in Figure 6.5.10, if the first Heads is immediately followed by Tails, our progress is destroyed and we must start from scratch. But this *is* progress for us in solving the problem, since the fact that the system can get reset suggests the strategy of first-step analysis. Let's condition on the outcome of the first toss:

$$E(W_{HH}) = E(W_{HH}|\text{first toss } H)\frac{1}{2} + E(W_{HH}|\text{first toss } T)\frac{1}{2}.$$

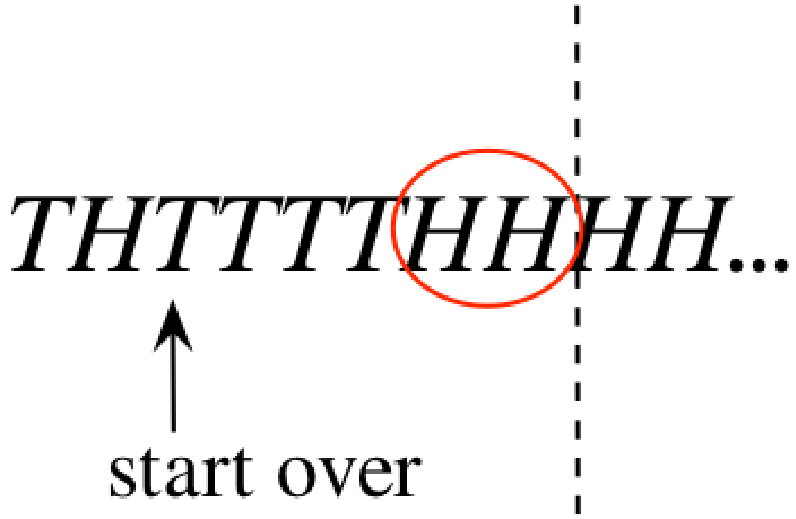

Figure 6.5.10: When waiting for *HH*, partial progress can easily be destroyed.

View Larger Image
Image Description

For the second term, $E(W_{HH}|\text{first toss T}) = 1 + E(W_{HH})$ by memorylessness. For the first term, we compute $E(W_{HH}|\text{1st toss H})$ by further conditioning on the outcome of the second toss. If the second toss is Heads, we have obtained *HH* in two tosses. If the second toss is Tails, we've wasted two tosses and have to start all over! This gives

$$E(W_{HH}|\text{first toss } H) = 2 \cdot \frac{1}{2} + (2 + E(W_{HH})) \cdot \frac{1}{2}.$$

Therefore,

$$E(W_{HH}) = \left(2 \cdot \frac{1}{2} + (2 + E(W_{HH})) \cdot \frac{1}{2}\right) \frac{1}{2} + (1 + E(W_{HH})) \frac{1}{2}.$$

Solving for $E(W_{HH})$, we get $E(W_{HH}) = 6$.

It might seem surprising at first that the expected waiting time for *HH* is greater than the expected waiting time for *HT*. How do we reconcile this with the fact that in two tosses of the coin, *HH* and *HT* both have a $1/4$ chance of appearing? Why aren't the average waiting times the same by symmetry?

As we solved this problem, we in fact noticed an important *asymmetry*. When waiting for *HT*, once we get the first Heads, we've achieved partial progress that cannot be destroyed: if the Heads is followed by another Heads, we're in the same position as before, and if the Heads is followed by a Tails, we're done. By contrast, when waiting for *HH*, even after getting the first Heads, we could be sent back to square one if the Heads is followed by a Tails. This suggests the average waiting time for *HH* should be longer. Symmetry implies that the average waiting time for *HH* is the same as that for *TT*, and that for *HT* is the same as that for *TH*, but it does not imply that the average waiting times for *HH* and *HT* are the same.

More intuition into what's going on can be obtained by considering a long string of coin flips, as in Figure 6.5.11. We notice right away that appearances of *HH* can overlap, while appearances of *HT* must be disjoint. Since there are the same average number of *HH*s and *HT*s, but *HH*s clump together while *HT*s do not, the spacing between successive strings of *HH*s must be greater to compensate.

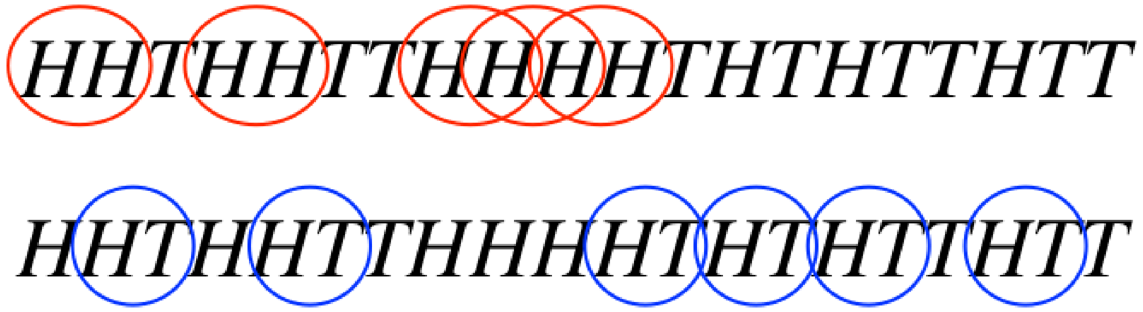

Figure 6.5.11: Clumping. (a) Appearances of *HH* can overlap. (b) Appearances of *HT* must be disjoint.

View Larger Image

Image Description

Related problems occur in information theory when compressing a message, and in genetics when looking for recurring patterns (called *motifs*) in DNA sequences.

Learn About Verified Certificates