6.5 Conditional Expectation

There are 2 closely linked notions of conditional expectation:

- Conditional expectation $E(Y|A)$ given an {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 an {random variable}: intuitively, $E(Y|X)$ is the r.v that best predicts Y using only information from X.

Definition 6.5.1 (Conditional expectation given an event)

Let A be event with positive probability. The conditional expectation of Y given A is:

- If Y is discrete r.v: $E(Y|A) = \sum_y y \cdot P(Y=y|A)$
- If Y is continuous r.v: $E(Y|A) = \int_{-\infty}^{\infty} y \cdot f(y|A)\, dy$,

where $f(y|A)$ is the derivative of CDF $F(y|A) = P(Y \leq y|A)$, and can be computed using Bayes' rule:

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

Warning 6.5.2 (Mistaken conditional and unconditional expectation)

Confusing conditional and unconditional expectation is dangerous. You should always be aware of what you are conditioning on.

Example 6.5.3 (Life expectancy)

Fred is 30 years old, his country life expectancy is 80. Should he conclude that, on average he has 50 years left to live?

No, because he missed 1 crucial fact: he has already lived to 30.

Let T be Fred's lifespan:

$$E(T) < E(T \mid T \geq 30)$$

- To prove this, consider:

$$E(T \mid T \geq 30) = \sum_{t=30}^{\infty} t \cdot P(T=t \mid T \geq 30)$$

$$= \sum_{t=30}^{\infty} t \cdot \frac{P(T=t)}{P(T \geq 30)}$$

$$= \frac{1}{P(T \geq 30)} \sum_{t=30}^{\infty} t \cdot P(T=t)$$

$\frac{1}{P(T \geq 30)} \geq 1$

$\sum_{t=30}^{\infty} t \cdot P(T=t)$ subset of $E(T) = \sum_{t=0}^{\infty} t \cdot P(T=t)$

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:

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

↳ Connection with LOTP through fundamental bridge:

Let $Y = I_B$ for an event B, then:

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

Example 6.5.5 (Two-envelope paradox)

There are 2 sealed envelopes, on of which has twice the amount of money to the other. Which one should you choose?

Let X and Y represents the amount in the left and right envelopes.

. There are 2 theories:

- By symmetry that $E(X) = E(Y)$, there is no reason that you should choose one over the other
- Suppose that the left has \$100, then the right envelope has either \$50 or \$200, which average to \$125. This contradict the first theory, and since we can apply the same reasoning to the right envelope, leading to switching back and forth between 2 envelopes.

Formalizing this, we have $Y = 2X$ or $Y = \frac{X}{2}$ with equal probabilities.

By LOTE:

$$E(Y) = E(Y \mid Y = 2X) \cdot \frac{1}{2} + E(Y \mid Y = \frac{X}{2}) \cdot \frac{1}{2}$$

One might think that:

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

which is 25% from switching (by theory 2).

<u>However</u>, there is no justification for $E(Y \mid Y = 2X) = E(2X)$ or $E(Y \mid Y = \frac{X}{2}) = E(\frac{X}{2})$. To see why, Let I indicate $Y = 2X$, then:

$$E(Y \mid 2X) = E(2X \mid \underbrace{I = 1}_{\text{we can only drop this term if } X \text{ and } I \text{ are independent, but they are not (definition of } I \text{ above)}})$$

<u>Example 6.5.7 (Mystery prize)</u>

You are bidding on a mystery box containing a mystery prize. The prize is unknown, except that its worth between 0 - 1 million dollars. So the true value V of the prize is Uniform on $[0, 1]$ (in million of dollars)

You can bid any amount b (in millions of dollars). Specifically, the bid is
- rejected if $b < \frac{2}{3}V$
- accepted if $b \geq \frac{2}{3}V$, your payoff is $V - b$

How much should you bid?

- Let W be the payoff amount.

To find the expected payoff amount, condition on whether the bid is accepted. The payoff is
- $V - b$ if the bid is accepted
- 0 if the bid is rejected

So by LOTE:

$$\begin{aligned} E(W) &= E\left(W \mid b \geq \tfrac{2}{3}V\right) \cdot P\left(b \geq \tfrac{2}{3}V\right) + E\left(W \mid b < \tfrac{2}{3}V\right) \cdot P\left(b < \tfrac{2}{3}V\right) \\ &= E\left(V - b \mid b \geq \tfrac{2}{3}V\right) \cdot P\left(b \geq \tfrac{2}{3}V\right) + \underbrace{E\left(0 \mid b < \tfrac{2}{3}V\right) \cdot P\left(b < \tfrac{2}{3}V\right)}_{\text{rejected case} = 0} \\ &= E\left(V - b \mid b \geq \tfrac{2}{3}V\right) \cdot P\left(b \geq \tfrac{2}{3}V\right) \\ &= \left(E\left(V \mid V \leq \tfrac{3}{2}b\right) - b\right) \cdot P\left(V \leq \tfrac{3}{2}b\right) \end{aligned}$$

- For $b \geq \frac{2}{3}$, $P\left(V \leq \frac{3}{2}b\right) = P(V \leq 1) = 1$ $\quad \langle V \sim \text{Uniform} \rangle$

So the RHS equals $\frac{1}{2} - b < 0 \Rightarrow$ you lose money

- For $b < \frac{2}{3}$, $P\left(V \leq \frac{3}{2}b\right) = \frac{3}{2}b$, so the conditional distribution of V is Uniform on $\left[0, \frac{3}{2}b\right]$.

So the RHS equals $\left(E\left(V \mid V \leq \frac{3}{2}b\right) - b\right) P\left(V \leq \frac{3}{2}b\right)$

$$= \left(\tfrac{3}{4}b - b\right)\left(\tfrac{3}{2}b\right)$$

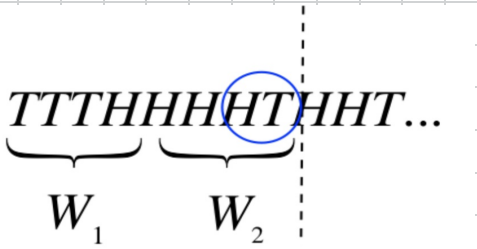
$$= -\tfrac{3}{8}b^2 \Rightarrow \text{You also lose money}$$

- In conclusion, you should not play this game.

Example 6.5.8 (Time until HH vs HT)

A fair coin is tossed repeatedly. What is the expected number of tosses until the pattern HT appears for the first time? What is the expected number of tosses until the pattern HH appears for the first time?

- Let W_{HT} be the number of tosses until HT appears.



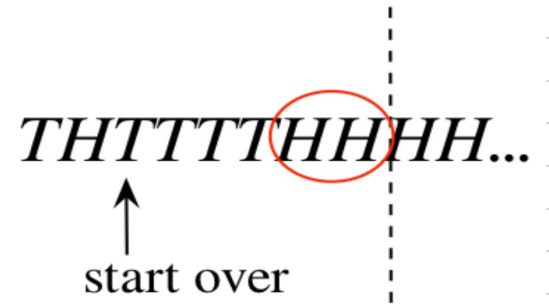
From the image we can see that $W_{HT} = W_1 + W_2$

so by linearity, $E(W_{HT}) = E(W_1) + E(W_2)$

And since $W_1, W_2 \overset{i.i.d}{\sim} FS(\frac{1}{2})$ (W_1 is tosses until first Head, W_2 is tosses until first Tail)

$\Rightarrow E(W_1) = E(W_2) = \frac{1}{p} = 2 \Rightarrow E(W_{HT}) = 4$

- Let W_{HH} be number of tosses until HH appears.



From the image if we toss
- Head, then we progress to the next stage and wait for another Head
- Tail, then the problem reset

The "reset" nature of the problem makes it perfect for solving by conditioning on the first step, so by LOTE:

$$E(W_{HH}) = \underbrace{E(W_{HH} \mid \text{first toss H}) \cdot \frac{1}{2}}_{\text{first term}} + \underbrace{E(W_{HH} \mid \text{first toss T}) \cdot \frac{1}{2}}_{\text{second term}}$$

- For second term, $E(W_{HH} \mid \text{first toss } T) = 1 + E(W_{HH})$ by memorylessness.
- For first term, $E(W_{HH} \mid \text{first toss } H)$ is computed by further conditioning on the second toss, so:

$$E(W_{HH} \mid \text{first toss } H) = E(W_{HH} \mid \text{first toss } H, \text{second toss } H) \cdot \frac{1}{2} + E(W_{HH} \mid \text{first toss } H, \text{second toss } T) \cdot \frac{1}{2}$$

$$= \underbrace{2 \cdot \frac{1}{2}}_{\text{get pattern } HH \text{ in 2 tosses}} + \underbrace{(2 + E(W_{HH})) \cdot \frac{1}{2}}_{\text{wasted 2 tosses and start over}}$$

- Replacing first and second terms in the original equation, gives:

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

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

Why $E(W_{HH}) = 6 > E(W_{HT}) = 4$ when probability of event HH and HT are both $1/4$?

- When waiting for HT, if H is tossed, we achieve partial progress where if we get another H, we are at the same position and if we get T, we are done
- When waiting for HH, if H is tossed and follows by T, we lose all our progress, this suggest average waiting time for HH is longer.
- Symmetry implies that average waiting time for HH and TT are the same, and HT and TH is the same, not HH and HT.

Definition 6.6.1 (Conditional Expectation given an random variable)

- Let $g(x) = E(Y \mid X = x)$. Then the conditional expectation of Y given X, denoted $E(Y \mid X)$, is defined to be random variable $g(X)$.
- In other words, if after doing experiment X crystallizes into x, then $E(Y \mid X)$ crystallizes into $g(x)$.
- If Y is discrete:

$$g(x) = E(Y \mid X = x) = \sum_y y \cdot P(Y = y \mid X = x)$$

- If Y is continuous:

$$g(x) = E(Y \mid X = x) = \int_{-\infty}^{\infty} y \cdot f_{Y|X}(y \mid x)\, dy$$

Warning 6.6.2 (Notion confusion)

The notion can sometimes cause confusion. It doesn't say:

$$g(x) = E(Y \mid X = x), \text{ so } g(X) = E(Y \mid X = X) = E(Y)$$

Rather, we should first compute function $g(x)$, then plug X for x. For example, if $g(x) = x^2$, then $g(X) = X^2$

Warning 6.6.3 (Difference between $E(Y|X)$ and $E(Y|A)$)

- By definition, $E(Y|X)$ is a function of X, so its a random variable. Thus, make sense to compute values like the mean $E(E(Y|X))$, and variance $Var(E(Y|X))$.
- Worth keeping in mind that:
  - $E(Y|A)$ are numbers
  - $E(Y|X)$ are random variables

Example 6.6.4 (Mean and variance of $E(Y|X)$)

Suppose we have a stick of length 1, we break it at point X chosen uniformly at random. Given that $X = x$, we then choose another break point Y uniformly on interval $[0, x]$

Find $E(Y|X)$, and its mean, variance.

- We have $X \sim \text{Unif}(0,1)$ and $Y|X=x \sim \text{Unif}(0,x)$, so:

$$E(Y|X=x) = \frac{x}{2}$$

Plugging X for x, we have:

$$E(Y|X) = \frac{X}{2}$$

- The expected value of $E(Y|X)$ is:

$$E(E(Y|X)) = E\left(\frac{X}{2}\right) = \frac{1}{4}$$

- The variance of $E(Y|X)$ is:

$$\text{Var}(E(Y|X)) = \text{Var}\left(\frac{X}{2}\right) = \frac{1}{48}$$

6.7 Adam's law and other properties of conditional expectation

Some useful properties of conditional expectation

Theorem 6.7.1 (Dropping what's independent)

If X and Y are independent, then $E(Y|X) = E(Y)$

→ Proof:

- Independence implies $E(Y|X=x) = E(Y)$ for all x, so $E(Y|X) = E(Y)$
- Intuitively, if X provides no information about Y, the our best guess for Y, even if we know X, is still $E(Y)$.

↓ The converse is false:

Counterexample is given in Example 6.7.4

Example 6.7.4 ($E(Y|X) = E(Y)$ doesn't mean X and Y are independent)

- Let $X_1, \dots, X_n$ be i.i.d, and $S_n = X_1 + \dots X_n$. Find $E(X_1|S_n)$
- By linearity: $E(X_1|S_n) + \dots + E(X_n|S_n) = E(S_n|S_n) = S_n$
- And by symmetry: $E(X_1|S_n) = \dots = E(X_n|S_n)$
- Therefore: $E(X_1|S_n) = \frac{S_n}{n} = \bar{X}_n$

which is the sample mean of all X_j's

Theorem 6.7.2 (Taking out what's known)

For any function h,

$$E(h(X) \cdot Y|X) = h(X) \cdot E(Y|X)$$

→ Proof:

Intuitively, when we take expectations condition on X, we are treating as if X crystallize into x. Then any function of X, like $h(X)$, also crystallize into $h(x)$.

Theorem 6.7.3 (Linearity)

$$E(Y_1 + Y_2 \mid X) = E(Y_1 \mid X) + E(Y_2 \mid X)$$

Proof:

This is arrived from the fact that: $E(Y_1 + Y_2) = E(Y_1) + E(Y_2)$

Theorem 6.7.4 (Adam's Law)

For any r.v.s X and Y,

$$E(E(Y \mid X)) = E(Y)$$

Proof:

This is proof for X and Y as discrete

let $E(Y \mid X) = g(X)$, we have

$$E(X) = \sum_x x \cdot P(X = x) \quad \langle \text{definition of expectation} \rangle$$

$$\Leftrightarrow \quad E(g(X)) = \sum_x g(x) \cdot P(X = x) \quad \langle \text{LOTUS} \rangle$$

$$= \sum_x \left(\sum_y y \cdot P(Y = y \mid X = x) \right) \cdot P(X = x) \quad \langle \text{definition of } g(x) \rangle$$

$$= \sum_x \sum_y y \; P(X = x) \cdot P(Y = y \mid X = x)$$

$$= \sum_x \sum_y y \; P(X = x, Y = y) \quad \langle \text{Bayes' rule} \rangle$$

$$= \sum_y y \cdot \sum_x P(X = x, Y = y) \quad \langle \text{Swapping sum} \rangle$$

$$= \sum_y y \cdot P(Y = y) \quad \langle \text{sum over all } x \rangle$$

$$= E(Y) \quad \langle \text{definition of expectation of } Y \rangle$$

- Apply the same logic for continuous case.

- Adam's law connects conditional expectation to unconditional expectation

Example 6.7.6 (Linear regression)

Linear regression predict Y based off X, it assumes the conditional expectation of Y is linear in X:

$$E(Y|X) = a + bX$$

a) Show that this is equivalent to: $Y = a + bX + \varepsilon$, where ε is a r.v (error) with $E(\varepsilon|X) = 0$.

We have: $Y = a + bX + \varepsilon$, where $E(\varepsilon|X) = 0$

$$\Leftrightarrow E(Y|X) = E(a|X) + E(bX|X) + E(\varepsilon|X) \quad \langle \text{linearity} \rangle$$

$$= E(a) + bX\,E(1|X) + 0 \quad \langle \text{properties of conditional} \rangle$$

$$= a + bX$$

b) Solve for constants a and b in terms of $E(X)$, $E(Y)$, $Cov(X,Y)$ and $Var(X)$

From a: $E(Y|X) = a + bX + \varepsilon$

$$\Leftrightarrow E(E(Y|X)) = a + bE(X) + E(\varepsilon) \quad \langle \text{Expectation both sides} \rangle$$

$$\Leftrightarrow E(Y) = a + bE(X) + 0 \quad \langle \text{LHS: Adam's Law, RHS: } E(\varepsilon) = E(E(\varepsilon|X)) = 0 \rangle$$

$$\Leftrightarrow a = E(Y) - bE(X)$$

Taking covariance with X of both sides in $Y = a + bX + \varepsilon$:

$$Cov(X,Y) = Cov(X,a) + bCov(X,X) + Cov(X,\varepsilon)$$

$$= b\,Var(X)$$

Thus:

$$\begin{cases} b = \dfrac{Cov(X,Y)}{Var(X)} \\ a = E(Y) - \dfrac{Cov(X,Y)}{Var(X)}E(X) \end{cases}$$

6.8 Eve's law and conditional variance

Definition 6.8.1 (Conditional Variance)

The conditional variance of Y given X is:

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

This is equivalent to:

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

→ Almost the same as unconditional variance:

We basically just replace all instances of $E(.)$ in the definition with $E(.|X)$

→ Connection with conditional expectation

Since conditional variance is defined by conditional expectations, we can use results about conditional expectations to calculate conditional variance.

Warning 6.8.2 (conditional variance is a random variable)

Like $E(Y|X)$, $\text{Var}(Y|X)$ is a random variable, and is a function of X.

Theorem 6.8.4 (Eve's Law) or Law of Total Variance

For any r.v.s X and Y,

$$\text{Var}(Y) = E(\text{Var}(Y|X)) + \text{Var}(E(Y|X))$$

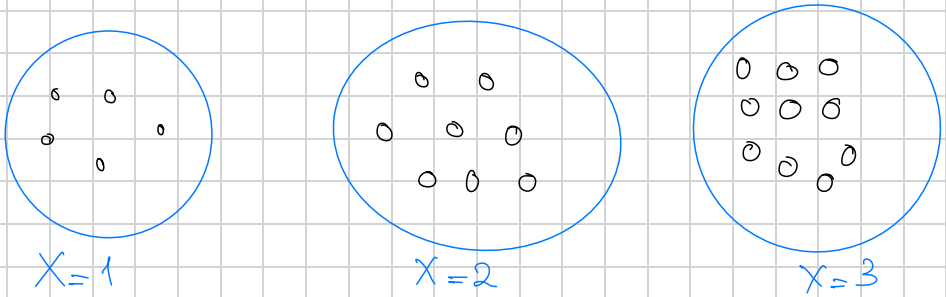
→ Proof: Let $g(X) = E(Y|X)$. By Adam's Law, $E(g(X)) = E(Y)$.

- $E(\text{Var}(Y|X)) = E(E(Y^2|X) - g(X)^2) = E(Y^2) - E(g(X)^2)$
- $\text{Var}(E(Y|X)) = E(g(X)^2) - (Eg(X))^2 = E(g(X)^2) - (EY)^2$

Adding these we have $\text{Var}(Y)$

Visualizing Eve's Law

We split sample space by r.v X, i.e:



Let Y be r.v represent number of pebbles

Eve's law says, in general:

total variance = within-group variance + between-group variance

$E(Y)$ (average pebbles within each groups) $E(Var(Y|X))$ + (variation of group means) $Var(E(Y|X))$

Warning 6.8.6 (Confusion with LOTE)

Its wrong to say that:

$$Var(Y) = Var(Y|A) \cdot P(A) + Var(Y|A^c) \cdot P(A^c)$$

Rather, we should use Eve's law, let I be indicator event A:

$$Var(Y) = E(Var(Y|I)) + Var(E(Y|I))$$

Example 6.8.7 (Random sum)

A store has N customers in a day, where N is r.v with finite mean and variance.

Let X_j be the amount customer j^{th} spend in that day, assume X_j has mean μ and variance σ^2

N and all of X_j are independent of each other.

Find the mean and variance of random sum $X = \sum_{j=1}^{N} X_j$, which is the store revenue in a day, in terms of μ, σ^2, $E(N)$ and $Var(N)$

- We might try to find mean by linearity: $E(X) = N\mu$

  But:
  - $E(X)$ is function of N and X_j, X is not a sum (number), but a random sum (function of N and X_j)
  - $N\mu$ is a random variable (function of N)

  And linearity only applies to fixed numbers (constants)

- So lets X to function of just N, by condition on N:

$$\begin{aligned} E(X|N) &= E\left(\sum_{j=1}^{N} X_j \,|\, N\right) && \text{< linearity of conditional expectation} \\ &= \sum_{j=1}^{N} E(X_j|N) && \text{< Taking out what's known} \\ &= \sum_{j=1}^{N} E(X_j) && \text{< } X_j \text{ and } N \text{ are independent} \\ &= N \cdot \mu \end{aligned}$$

- "$E(X|N) = N\mu$" is valid since both sides of the equation are random variables (functions of N)

- Now we can find $E(X)$ by Adam's Law:

$$E(X) = E\left(E(X|N)\right) = E(N\mu) = \mu E(N)$$

This means:

average revenue = average spend per customer x average number of customers

- To find $Var(X)$, we condition on N to turn $Var(X)$ from function of N and X_j (since X is function of N and X_j) to function of just N.

$$\begin{aligned} Var(X|N) &= Var\left(\sum_{j=1}^{N} X_j \,|\, N\right) \\ &= \sum_{j=1}^{N} Var(X_j|N) && \text{< taking out what's known} \\ &= \sum_{j=1}^{N} Var(X_j) = N\sigma^2 && \text{< } X_j \text{ and } N \text{ independent} \end{aligned}$$

- Then use Eve's law to get $Var(X)$:

$$\begin{aligned} Var(X) &= E(Var(X|N)) + Var(E(X|N)) \\ &= E(N\sigma^2) + Var(N\mu) \\ &= \sigma^2 E(N) + \mu Var(N) \end{aligned}$$