Lecture 11:

Conditional expectation

Conditional expectation

Conditional expectation is, basically, expectation but computed with conditional probabilities.
There are 2 different but closely related notions of c.e:

1. C.e. $E(Y|A)$ **given an event**: Y is a r.v., A is an event. If we learn that A occurred, our updated expectation for Y is denoted by $E(Y|A)$ and is computed analogously to $E(Y)$, except using **conditional** probabilities given A.

2. C.e. $E(Y|X)$ **given a random variable**: is is a subtle question – how do define $E(Y|X)$ where both X and Y are r.v.-s. Intuitively, $E(Y|X)$ is the r.v. that best predicts Y using only the information available from X.

Conditional expectation given an event

Conditional expectation given an event

Definition 9.1.1 (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 $E(Y|A) = \sum_y y\,P(Y = y\,|A)$, where the sum is over Supp(Y).

If Y is a continuous r.v. with PDF f, then $E(Y|A) = \int_{-\infty}^{+\infty} y\,f(y\,|A)\,dy$

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

Conditional expectation given an event

Intuition 9.1.2 Imagine running an experiment n times, then the expectation of Y is the limit of: $E(Y) \approx \frac{1}{n}\sum_{j=1}^{n} y_j$

To approximate $E(Y \mid A)$, we only take the runs where A occurred, and average only **those** Y-values: $E(Y \mid A) \approx \frac{\sum_{j=1}^{n} y_j \, I_j(A)}{\sum_{j=1}^{n} I_j(A)}$

where $I_j(A)$ is the indicator of A happening on j-th run.

Conditional expectation given an event

LOTP allows us to get unconditional probabilities by slicing up the sample space and computing conditional probabilities in each slice. Same idea works here:

Theorem 9.1.5 (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 r.v. on this sample space. Then $E(Y) = \sum_{i=1}^{n} E(Y \mid A_i)\, P(A_i)$

In fact, since all probabilities are expectations (fundamental bridge), LOTP is a special case of this (LOTE). To see this, let $Y = I_B$ for an event B, then the above says

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

Conditional expectation given an event

One should be careful with conditioning expectations:

Example 9.1.6 (Two-envelope paradox). You are given two envelopes, each with some positive amount of money. You are told that one of the envelopes contains twice as much as the other. Which one do you prefer – 1st or 2nd?

Denote the sums X, Y. So either $Y = 2X$ or $Y = X/2$ with equal probabilities. By symmetry, there is no reason to prefer one over the other – $E(X) = E(Y)$.

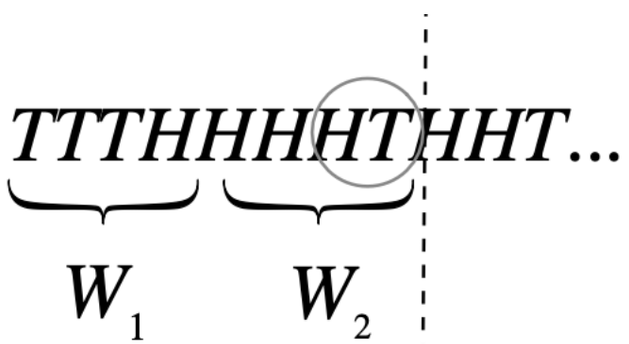
But suppose $X = x$ – then Y is either $2x$ or $x/2$:

$E(Y) = E(Y \mid Y = X) \cdot 1/2 + E(Y \mid Y = X/2) \cdot 1/2$ – one might think that this is $E(2X) \cdot 1/2 + E(X/2) \cdot 1/2 = 5/4\, E(X)$

The resolution is that X and indicator of, say, $Y = 2X$ – are ***not independent***! – somehow, observing X gives info whether its bigger

Conditional expectation given an event

Example 9.1.9 (Time until HH vs HT). You toss a fair coin repeatedly. What is the expected # of tosses till you first see HT? HH?



<- wait W_1 for first H, then wait W_2 for first T after that. W_1, W_2 have First Success dist, $FS(1/2)$, so $E(W_1) = E(W_2) = 2$

and so $E(W_{HT}) = 4$

It's not so simple for HH! First, we have

$$E(W_{HH}) = E(W_{HH} \mid 1\text{st was H})\frac{1}{2} + E(W_{HH} \mid 1\text{st was T})\frac{1}{2}$$

For the latter, $E(W_{HH} \mid 1\text{st was T}) = 1 + E(W_{HH})$. For the former – we either get H on second toss, or we've wasted 2 tosses:

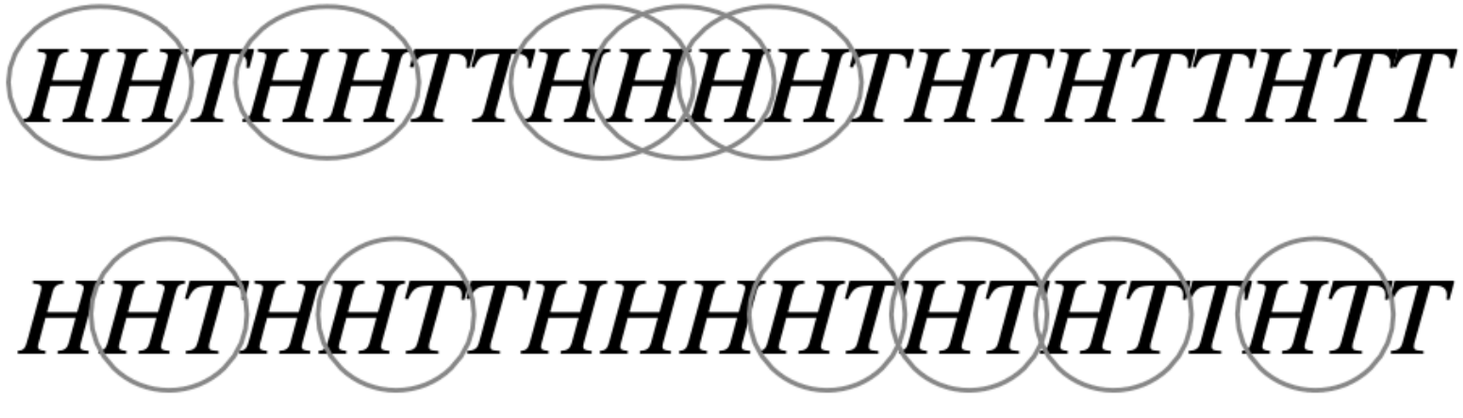
$E(W_{HH} \mid 1\text{st was H}) = 2\frac{1}{2} + (2 + E(W_{HH}))\frac{1}{2}$. From that, $E(W_{HH}) = 6$

Conditional expectation given an event

So we got $E(W_{HT}) = 4$, but $E(W_{HH}) = 6$ – why such an asymmetry?

One explanation is: waiting for HT, if we rolled H, we've made some progress: if we get H again, we only need to wait for another T again. Waiting for HH, if we rolled the first H, we could be sent back to square one, if we get a T after it.

Another explanation is that appearances of HH overlap, but average # of HH-s and HT-s is the same, so average waiting time for HH-s must be larger to compensate for that:



Conditional expectation given an r.v.

Conditional expectation given an r.v.

So we know what $E(Y|X=x)$ is (since $X=x$ is an event) – to compute it, we sum/integrate Y "out", so it is a function of x only. We can call it $g(x)=E(Y|X=x)$.

Example 9.2.4 A stick of length 1 is broken at point X chosen uniformly at random. Given that $X=x$, we choose another breakpoint Y uniformly at the interval $[0,x]$. Find $E(Y|X)$, its mean and variance.

So $X \sim$ Unif$(0,1)$ and $(Y|X=x) \sim$ Unif$(0,x)$. Then $E(Y|X=x)=x/2$, so, **plugging in X for x** – $E(Y|X)=X/2$.
The expected value: $E(E(Y|X))=E(X/2)=1/4$
The variance: Var$(E(Y|X))=$ Var$(X/2)=1/48$

Properties of conditional expectation

Properties of conditional expectation

- Dropping what's independent:

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

- Taking out what's known:

  For any function h, $E(h(X)Y|X) = h(X)\,E(Y|X)$

- Linearity:

  $E(Y_1 + Y_2|X) = E(Y_1|X) + E(Y_2|X)$ and $E(cY|X) = c\,E(Y|X)$

- Adam's law: $E(E(Y|X)) = E(Y)$

- Projection interpretation:

  The r.v. $Y - E(Y|X)$ – aka the **residual**, from using X to predict Y, is uncorrelated with $h(X)$ for any function h.

Properties of conditional expectation

Theorem 9.3.2 (Taking out what's known). For any function h, $E(h(X)Y|X) = h(X)\,E(Y|X)$.

Proof idea: This is because, when taking expectation given X, we are treating X as if it has crystallised into a known constant – so we can take out $h(X)$ as a constant.

Theorem 9.3.7 (Adam's law). For any r.v.-s X, Y:

$E(E(Y|X)) = E(Y)$.

Proof: LOTUS for $E(Y|X) = g(X)$:

$$E(g(X)) = \Sigma_x\, g(x)\, P(X = x) = \Sigma_x \left(\Sigma_y\, yP(Y = y|X = x)\right)P(X = x) =$$

$$= \Sigma_x \Sigma_y\, y\, P(X = x)\, P(Y = y|X = x) =$$

$$= \Sigma_y\, y \Sigma_x\, P(X = x, Y = y) = \Sigma_y\, y\, P(Y = y) = E(Y)$$

Properties of conditional expectation

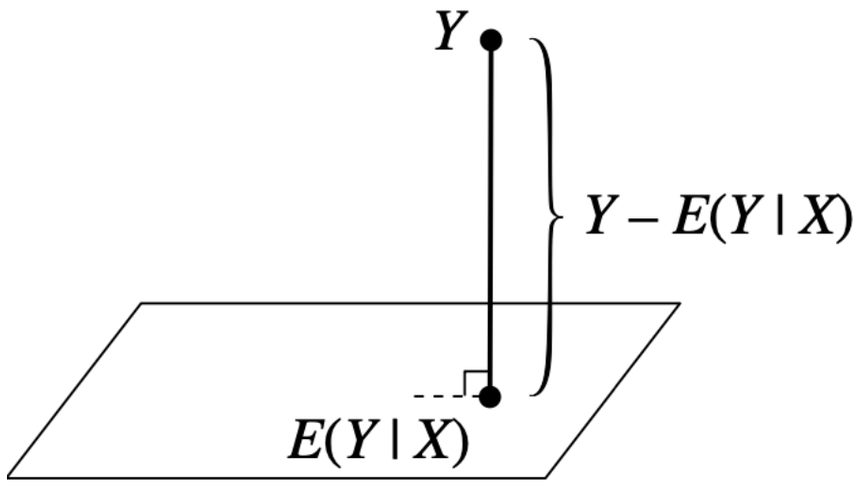
Theorem 9.3.9 (Projection interpretation). For any function h, the r.v. $Y - E(Y|X)$ is uncorrelated with $h(X)$.

Equivalently, $E((Y - E(Y|X))h(X)) = 0$

Proof: We have

$$E((Y - E(Y|X))h(X)) = E(h(X)Y) - E(h(X)E(Y|X)) =$$
$$= E(h(X)Y) - E(E(h(X)Y|X)).$$

By Adam's law, the latter term $= E(h(X)Y)$.



<- the plane = "space of all functions of X"

the residual $Y - E(Y|X)$ is "orthogonal" (uncorrelated) with any function of X