Conditional Expectations

David S. Rosenberg

NYU: CDS

January 27, 2021

Goal of this lecture

- This class has a lot of conditional expectation calculations.
- We assume that you've seen these concepts in probability classes.
- Goal for this lecture: [re]building your fluency with these calculations.

Keeping things simple

- For any random element $X \in \mathcal{X}$ we consider,
  - We'll assume $|\mathcal{X}| < \infty$.
  - That is, assume X can only take finitely many possible values.
- Then distribution of X is represented by its **probability mass function (PMF)**
- All the results generalize, but definitions get more complicated.
- Remember that the point is to give you practice in applying the theorems to do calculations.

Contents

Basic expectations

Random elements vs. random variables

- The generic term for something that's random is **random element**.
- The specific term for a **real-valued** random element is **random variable**.
- We'll only be talking about expectations of random variables.

Basic expectation

- Let $Y \in \mathcal{Y} \subset \mathbb{R}$ be a random variable with PMF $p(y)$.
- For simplicity, we'll assume $\mathcal{Y}$ is finite.
- Then the **expectation of** Y is defined as

$$\mathbb{E}Y = \sum_{y \in \mathcal{Y}} y p(y).$$

We write expectations of r.v.'s, but it's best to think of expectations as properties of distributions.

Expectation of $f(X)$

- Let $X \in \mathcal{X}$ be a random element.
- Let $f : \mathcal{X} \to \mathbb{R}$ be an ordinary real-valued function.
- Then $Y = f(X)$ is a random variable.
- The expectation of $f(X)$ is

$$\mathbb{E}f(X) = \sum_{x \in \mathcal{X}} f(x)p(x)$$

- We can derive this from our definition of expectation.

Conditional expectations

Conditional distributions

- Let $X \in \mathcal{X}$ be a random element.
- Let $Y \in \mathcal{Y} \subset \mathbb{R}$ be a random variable (r.v.)
- Let X, Y have joint PMF $p(x, y)$.
- The **conditional distribution of** Y **given** $X = x$ is given by the conditional PMF

$$p(y \mid x) = \frac{p(x, y)}{p(x)}.$$

- For each fixed x, $p(y \mid x)$ gives a distribution over $y \in \mathcal{Y}$.
- You can verify that for each $x \in \mathcal{X}$, $\sum_{y \in \mathcal{Y}} p(x, y) = 1$ and $p(x, y) \in [0, 1]$.

$\mathbb{E}[Y \mid X = x]$

Definition

The **conditional expectation of** Y **given** $X = x$, is the expectation of the distribution represented by $p(y \mid x)$. That is,

$$\mathbb{E}[Y \mid X = x] = \sum_{y \in \mathcal{Y}} y p(y \mid x).$$

$\mathbb{E}[Y \mid X]$

- $\mathbb{E}[Y \mid X = x]$ is an ordinary function of $x \in \mathcal{X}$. (Nothing random)
- To emphasize this, we can define $f(x) := \mathbb{E}[Y \mid X = x]$.
- We can now define $\mathbb{E}[Y \mid X]$:

Definition

We define the **conditional expectation of Y given X** as

$$\mathbb{E}[Y \mid X] = f(X),$$

where $f(x) := \mathbb{E}[Y \mid X = x]$.

- Since X is random, $f(X)$ and thus $\mathbb{E}[Y \mid X]$ are random variables.

Exercise

Show that $\mathbb{E}[h(X)\mathbb{E}[Y \mid X]] = \sum_{x \in \mathcal{X}} p(x)h(x)\mathbb{E}[Y \mid X = x]$.

Proof.

Let $f(x) = \mathbb{E}[Y \mid X = x]$. Then

$$
\begin{aligned}
\mathbb{E}[h(X)\mathbb{E}[Y \mid X]] &= \mathbb{E}[h(X)f(X)] \\
&= \sum_{x \in \mathcal{X}} p(x)h(x)f(x) \\
&= \sum_{x \in \mathcal{X}} p(x)h(x)\mathbb{E}[Y \mid X = x].
\end{aligned}
$$

$\square$

Identities for conditional expectations

Basic identities

- **Independence:** $\mathbb{E}[Y \mid X] = \mathbb{E}[Y]$ if X and Y are independent.
- **Taking out what is known:** $\mathbb{E}[h(X)Z \mid X] = h(X)\mathbb{E}[Z \mid X]$.
    - Generalization of $\mathbb{E}[cZ] = c\mathbb{E}Z$.
- **Linearity**: $\mathbb{E}[aX + bY \mid Z] = a\mathbb{E}[X \mid Z] + b\mathbb{E}[Y \mid Z]$, for any $a, b \in \mathbb{R}$.

Exercise

Show $\mathbb{E}[f(Z)X + g(Z)Y \mid Z] = f(Z)\mathbb{E}[X \mid Z] + g(Z)\mathbb{E}[Y \mid Z]$, for any $f, g : \mathcal{Z} \to \mathbb{R}$.

Proof.

We have

$$\begin{aligned} & \mathbb{E}[f(Z)X + g(Z)Y \mid Z] \\ &= \mathbb{E}[f(Z)X \mid Z] + \mathbb{E}[g(Z)Y \mid Z] \quad \text{linearity} \\ &= f(Z)\mathbb{E}[X \mid Z] + g(Z)\mathbb{E}[Y \mid Z] \quad \text{taking out what is known.} \end{aligned}$$

$\square$

Adam's Law / Law of Iterated Expectation

- $\mathbb{E}[Y \mid X]$ is a rv. What is its expectation?
- **Adam's Law**: $\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}Y$.
- Let $f(x) = \mathbb{E}[Y \mid X = x]$. So $f(X) = \mathbb{E}[Y \mid X]$ (by definition) and

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[Y \mid X]] &= \mathbb{E}[f(X)] \\
&= \sum_{x \in \mathcal{X}} p(x) f(x) \\
&= \sum_{x \in \mathcal{X}} p(x) \mathbb{E}[Y \mid X = x].
\end{aligned}$$

- So $\mathbb{E}Y$ can be computed as a weighted average of $\mathbb{E}[Y \mid X = x]$.

Proof of Adam's Law

- We have

$$
\begin{aligned}
\mathbb{E}[\mathbb{E}[Y \mid X]] &= \sum_{x\in\mathcal{X}} p(x)\mathbb{E}[Y \mid X = x] \quad \text{prev exercise} \\
&= \sum_{x\in\mathcal{X}} p(x)\left[\sum_{y\in\mathcal{Y}} yp(y \mid x)\right] \quad \text{def of cond exp} \\
&= \sum_{y\in\mathcal{Y}} y\left[\sum_{x\in\mathcal{X}} p(y \mid x)p(x)\right] \\
&= \sum_{y\in\mathcal{Y}} yp(y) \quad \text{Law of total probability} \\
&= \mathbb{E}Y
\end{aligned}
$$

Exercise (Partial expansion of expectation)

- Show that
$$\mathbb{E}[h(X)Y] = \sum_{x\in\mathcal{X}} p(x)h(x)\mathbb{E}[Y \mid X = x].$$
- A full expansion of the expectation would be a double sum over x and y:
$$\mathbb{E}[h(X)Y] = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} h(x)yp(x,y).$$
- With a single summation, the other sum is absorbed in $\mathbb{E}[Y \mid X = x]$.

Solution (Partial expansion of expectation)

- Let $f(x) = \mathbb{E}[Y \mid X = x]$. Then we have

$$\begin{aligned}
\mathbb{E}[h(X)Y] &= \mathbb{E}[\mathbb{E}[h(X)Y \mid X]] \quad \text{by Adam's Law} \\
&= \mathbb{E}[h(X)\mathbb{E}[Y \mid X]] \quad \text{taking out what is known} \\
&= \mathbb{E}[h(X)f(X)] \quad \text{definition} \\
&= \sum_{x \in \mathcal{X}} p(x)[h(x)f(x)] \quad \text{expectation of function} \\
&= \sum_{x \in \mathcal{X}} p(x)h(x)\mathbb{E}[Y \mid X = x] \quad \text{def of } f(x)
\end{aligned}$$

- Doing Adam's law followed by "taking out what is known" will be used for the majority of our calculations!

Exercise

- Recall the indicator function notation:

$$\mathbb{1}[W=1]=\begin{cases}1 & \text{if } W=1\\ 0 & \text{otherwise.}\end{cases}$$

- Show that

$$\mathbb{E}[\mathbb{1}[W=1]Y]=\mathbb{P}(W=1)\mathbb{E}[Y\mid W=1].$$

- You can either apply the previous exercise, or repeat the steps of the previous exercise.

Exercise solution

Proof.

Let $Z = \mathbb{1}[W = 1]$. Then

$$\begin{aligned}
\mathbb{E}[\mathbb{1}[W=1]Y] &= \mathbb{E}[\mathbb{E}(ZY \mid Z)] && \text{by Adam's Law} \\
&= \mathbb{E}[Z\mathbb{E}[Y \mid Z]] && \text{taking out what is known} \\
&= \mathbb{P}(Z=1)\cdot 1 \cdot \mathbb{E}[Y \mid Z=1] \\
&\quad + \mathbb{P}(Z=0)\cdot 0 \cdot \mathbb{E}[Y \mid Z=0] && \text{def of expectation} \\
&= \mathbb{P}(W=1)\mathbb{E}[Y \mid W=1] && \text{def of } Z
\end{aligned}$$

$\square$

Exercise: keeping just what is needed

- (1) Show that

$$\mathbb{E}[XY] = \mathbb{E}[X\mathbb{E}[Y \mid X]].$$

- For computing $\mathbb{E}[XY]$, we only care about the randomness in Y that is predictable by X.
  - Recall that $\mathbb{E}[Y \mid X] = f(X)$ is a deterministic function of X.
- (2) Show that

$$\mathbb{E}[h(X)Y] = \mathbb{E}[h(X)\mathbb{E}[Y \mid X]]$$

- Hint: Adam's Law followed by taking out what is known will work for each
- Note that (1) is a special case of (2), and you can also show (2) by combining 2 earlier exercises.

Projection interpretation

Inner product space of random variables

- Consider the space of all r.v.'s with finite variance.
- Give this space an inner product as follows:

$$\langle X, Y \rangle = \mathbb{E}[XY]$$

- The norm for this space is $\|X\| = \sqrt{\langle X, X \rangle} = \sqrt{\mathbb{E}X^2}$.
- The induced metric on this space is $d(X, Y) = \|X - Y\| = \sqrt{\mathbb{E}(X - Y)^2}$.
- This metric assesses how well one r.v. approximates another (in MSE)

Projections for random variables

Definition

Random variable S' is a **projection** of Y onto a set $\mathcal{S}$ of random variables if $S' \in \mathcal{S}$ and

$$\mathbb{E}\left(Y - S'\right)^2 \leq \mathbb{E}(Y - S)^2 \quad \forall S \in \mathcal{S}.$$

- In words, S' is the best approximation of Y in $\mathcal{S}$ in terms of mean squared error (MSE).
- We'll show that $\mathbb{E}[Y \mid X]$ is a projection of Y onto $\{h(X) \mid h \text{ is any real-valued function}\}$.

The residual

- We will think of $\mathbb{E}[Y \mid X]$ as an approximation to Y.
- And we will call $Y - \mathbb{E}[Y \mid X]$ the **residual** for the approximation.
- A residual is orthogonal to everything in the set we project onto.
- We next prove this property for $\mathbb{E}[Y \mid X]$... That is, we'll prove that

$$\langle Y - \mathbb{E}[Y \mid X], h(X) \rangle = 0 \quad \forall h : \mathcal{X} \to \mathbb{R}$$

- In terms of our specific inner product, we'll be showing that

$$\mathbb{E}[(Y - \mathbb{E}[Y \mid X])\, h(X)] = 0 \quad \forall h : \mathcal{X} \to \mathbb{R}$$

Projection interpretation theorem

Theorem (Projection interpretation)

For any $h : \mathcal{X} \to \mathbb{R}$, $\mathbb{E}[(Y - \mathbb{E}[Y \mid X])\, h(X)] = 0$.

Proof.

We have

$$
\begin{aligned}
&\mathbb{E}[(Y - \mathbb{E}[Y \mid X]) h(X)] \\
&\quad = \mathbb{E}[Yh(X)] - \mathbb{E}[\mathbb{E}[Y \mid X] h(X)] && \text{by linearity} \\
&\quad = \mathbb{E}[Yh(X)] - \mathbb{E}[\mathbb{E}[Yh(X) \mid X]] && \text{taking out what is known (in reverse)} \\
&\quad = \mathbb{E}[Yh(X)] - \mathbb{E}[Yh(X)] && \text{Adam's Law} \\
&\quad = 0
\end{aligned}
$$

$\square$

Orthogonality and correlation

Definition

The **covariance** of random variables X and Y is defined by

$$\mathrm{Cov}(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y) = \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y.$$

Definition

If $\mathrm{Cov}(X, Y) = 0$, then we say X and Y are **uncorrelated**.

Theorem

If X and Y are orthogonal (i.e. $\mathbb{E}[XY] = 0$), and $\mathbb{E}X = 0$, then $Cov(X, Y) = 0$.

Corollary

The residual $Y - \mathbb{E}[Y \mid X]$ and $h(X)$ are uncorrelated for every $h : \mathcal{X} \to \mathbb{R}$.

$\mathbb{E}[Y \mid X]$ gives the best prediction in MSE

Theorem (Conditional expectation minimizes MSE)

For random $X \in \mathcal{X}$ and $Y \in \mathbb{R}$, let $g(x) = \mathbb{E}[Y \mid X = x]$. Then

$$g(x) = \arg\min_{f} \mathbb{E}(Y - f(X))^2 .$$

Proof: $\mathbb{E}[Y \mid X]$ gives best prediction MSE

We have

$$
\begin{aligned}
&\mathbb{E}\left[(f(X)-Y)^2\right] \\
&= \mathbb{E}\left[f(X)-\mathbb{E}[Y \mid X]+\mathbb{E}[Y \mid X]-Y\right]^2 \\
&= \mathbb{E}(f(X)-\mathbb{E}[Y|X])^2+\mathbb{E}\left[(\mathbb{E}[Y|X]-Y)^2\right] \\
&\quad \underbrace{+2\,\mathbb{E}\left[\left(\underbrace{f(X)-\mathbb{E}[Y \mid X]}_{\text{function of } X}\right)\left(\underbrace{\mathbb{E}[Y \mid X]-Y}_{\text{residual}}\right)\right]}_{=0} \\
&= \mathbb{E}(f(X)-\mathbb{E}[Y|X])^2+\mathbb{E}\left[(\mathbb{E}[Y|X]-Y)^2\right] \quad \text{Projection interpretation}
\end{aligned}
$$

First term minimized by taking $f(x)=\mathbb{E}[Y \mid X=x]$. Second term is independent of f.

First variance decomposition

A decomposition with the residual

- Sometimes it's helpful to write Y as

$$Y = \underbrace{\mathbb{E}[Y \mid X]}_{\text{best prediction for } Y \text{ given } X} + \underbrace{Y - \mathbb{E}[Y \mid X]}_{\text{residual}}.$$

- From projection interpretation, $Y - \mathbb{E}[Y \mid X]$ is uncorrelated with any function of X.
- $\mathbb{E}[Y \mid X]$ is a function of X.
- If X and Y are uncorrelated r.v.'s, then

$$\operatorname{Var}(X + Y) = \operatorname{Var}(X) + \operatorname{Var}(Y).$$

- What can we do with this assortment of facts?

Variance decomposition with projection

Theorem (Variance decomposition with projection)

For any random $X \in \mathcal{X}$ and $Y \in \mathbb{R}$, we have

$$\operatorname{Var}(Y) = \operatorname{Var}(Y - \mathbb{E}[Y \mid X]) + \operatorname{Var}(\mathbb{E}[Y \mid X]).$$

- This imples $\operatorname{Var}(\mathbb{E}[Y \mid X]) \le \operatorname{Var}(Y)$, since variance is always ≥ 0.
- We can think of $\mathbb{E}[Y \mid X]$ as a “less random” version of Y.
- $\mathbb{E}[Y \mid X]$ only has the randomness in Y that is predictable from X. (why?)
- $\mathbb{E}[Y \mid X]$ is a deterministic function of X, so there's no other source of randomness in $\mathbb{E}[Y \mid X]$ than the randomness in X.

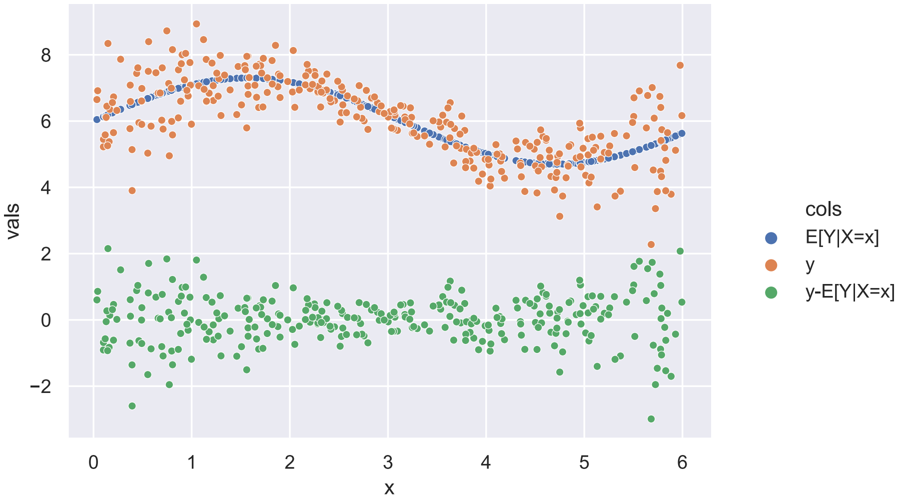
Empirical example of the variance decomposition

- Consider the following joint distribution of (X, Y):

$$\begin{aligned} X &\sim \text{Unif}[0,6] \\ Y \mid X = x &\sim \mathcal{N}\left(6 + 1.3\sin(x), \left[.3 + \frac{1}{4}|3 - x|\right]^2\right) \end{aligned}$$

- Given $X = x$, what's the best prediction for Y in MSE?
- It's $\mathbb{E}[Y \mid X = x] = 6 + 1.3\sin(x)$.

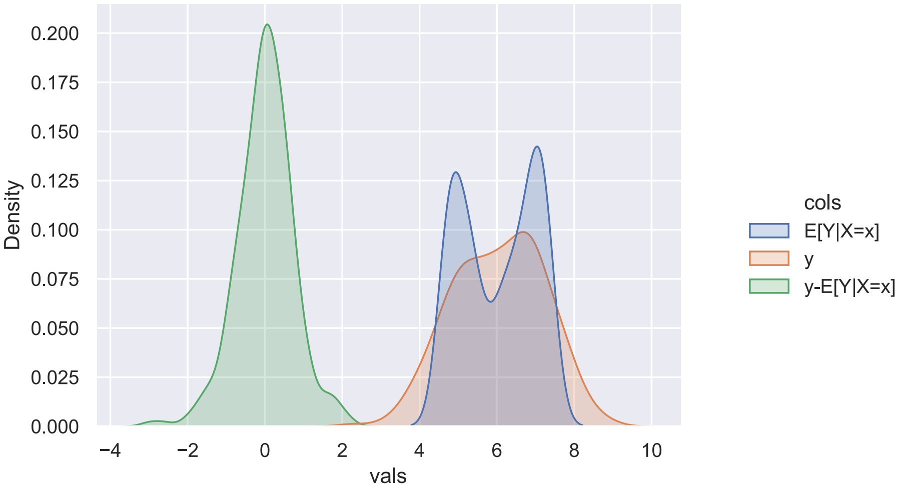
Draws from distribution



The graph shows a sample of size $n = 300$ from this distribution. For each sampled point (x, y), we also plot $(x, \mathbb{E}[Y \mid X = x])$, which is the best prediction of Y given $X = x$, along with the residual of that prediction. Note that the residuals hover around 0. Indeed, we should expect that since for any particular x, the conditional distribution of $Y \mid X = x$ has mean $\mathbb{E}[Y \mid X = x]$, which is exactly what we're subtracting off from Y in the residual. We can also compute this as follows:

$$
\begin{aligned}
&\mathbb{E}[Y - \mathbb{E}[Y \mid X] \mid X = x] \\
&\quad = \quad \mathbb{E}[Y \mid X = x] - \mathbb{E}[\mathbb{E}[Y \mid X] \mid X = x] \quad \text{by linearity} \\
&\quad = \quad \mathbb{E}[Y \mid X = x] - \mathbb{E}[Y \mid X = x]\mathbb{E}[1 \mid X = x] \quad \text{taking out what is known} \\
&\quad = \quad 0.
\end{aligned}
$$

Variance decomposition visualized



Variance decomposition estimtes

- By theorem: $\mathrm{Var}(Y) = \mathrm{Var}(Y - \mathbb{E}[Y \mid X]) + \mathrm{Var}(\mathbb{E}[Y \mid X])$.
- $\widehat{\mathrm{Var}}(Y - \mathbb{E}[Y \mid X]) \approx 0.53$
- $\widehat{\mathrm{Var}}(\mathbb{E}[Y \mid X]) \approx 0.91$
- $\widehat{\mathrm{Var}}(Y - \mathbb{E}[Y \mid X]) + \widehat{\mathrm{Var}}(\mathbb{E}[Y \mid X]) = 1.43$
- While $\widehat{\mathrm{Var}}(Y) \approx 1.39$.
- The gap between 1.43 and 1.39 is attributable to sampling error and vanishes as $n \to \infty$.

Conditional variance

Conditional variance

- Could take same approach as for conditional expectation:
  - Write $\text{Var}(Y \mid X = x)$ for the variance of the conditional distribution $Y \mid X = x$.
  - Let $f(x) = \text{Var}(Y \mid X = x)$
  - Then define $\text{Var}(Y \mid X) = f(X)$. Note that this is a random variable via X.
- Equivalently, we can just use conditional expectatons in the definition:

Definition

The **conditional variance** of Y given X is

$$\begin{aligned}\text{Var}(Y \mid X) &= \mathbb{E}\left[(Y - \mathbb{E}[Y \mid X])^2 \mid X\right] \\ &= \mathbb{E}\left[Y^2 \mid X\right] - (\mathbb{E}[Y \mid X])^2.\end{aligned}$$

Law of total variance / Eve's law

Also known as the variance decomposition formula,
the conditional variance forumula, and the law of iterated variances...

Theorem (Eve's Law)

For any random $X \in \mathcal{X}$ and $Y \in \mathbb{R}$,

$$Var(Y) = \mathbb{E}[Var(Y \mid X)] + Var(\mathbb{E}[Y \mid X]).$$

- If we write E for expectation and V for variance, the sequence of operations is EVVE.
- That's why this is sometimes called "Eve's law".
- This must also be why Adam's Law is called Adam's Law.

Exercise: Prove this by expanding both terms on the RHS and using Adam's Law.

Reference

Resources

- Chapter 9 of Blitzstein and Hwang's *Introduction to Probability, Second Edition* is highly recommended for what we need to know about conditional probabilities [KBH19].
- It usually takes a while to build up to a full measure-theoretic treatment of conditional probability, but if you want to go that direction, I like David Williams's *Probability with Martingales*, though there are plenty of other options.

References I

[KBH19] Joseph K. Blitzstein and Jessica Hwang, *Introduction to probability second edition*, 2nd ed., Chapman and Hall/CRC, 2019.