Lecture 10: Conditional Expectation

Mathematical Statistics I, MATH 60061/70061

Thursday September 30, 2021

Reference: Casella & Berger, 4.4

Conditional expectation

- Conditional expectation $E(Y \mid A)$ given an event
  - Let Y be a random variable, and A be an event.
  - Given that A occurred, the updated expectation for Y is denoted by $E(Y \mid A)$.
- Conditional expectation $E(Y \mid X)$ given a random variable
  - Both X and Y are random variables.
  - $E(Y \mid X)$ is the random variable that best predicts Y using only the information available from X.

Conditional expectation given an event

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

$$E(Y \mid A) = \sum_y y P(Y = y \mid A),$$

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

$$E(Y \mid A) = \int_y y f(y \mid A) dy,$$

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

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

Geometric expectation redux

Let $X \sim \mathrm{Geom}(p)$. Interpret X as the number of Tails before the first Heads in a sequence of coin flips with probability p of Heads.

To get $E(X)$, we condition on the outcome of the first toss:

- If it lands Heads, then X is 0 and we are done.
- If it lands Tails, then we've wasted one toss and are back to where we started, by memorylessness.

Therefore,

$$\begin{aligned} E(X) &= E(X \mid \text{first toss } H) \cdot p + E(X \mid \text{first toss } T) \cdot q \\ &= 0 \cdot p + (1 + E(X)) \cdot q, \end{aligned}$$

which gives $E(X) = q/p$.

Conditional expectation given a random variable

Let $g(x) = E(Y \mid X = x)$. Then the conditional expectation of Y given X, $E(Y \mid X)$, is defined to be the *random variable* $g(X)$.

The key to understanding $E(Y \mid X)$ is first to understand $E(Y \mid X = x)$, the conditional expectation of Y given the *event* $X = x$:

- If Y is discrete,

$$E(Y \mid X = x) = \sum_y yP(Y = y \mid X = x).$$

- If Y is continuous,

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

Note that $E(Y \mid X = x)$ is a function of x.

Example: breaking a stick

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

Example: breaking a stick

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

From the description of the experiment, $X \sim \mathrm{Unif}(0, 1)$ and $Y \mid X = x \sim \mathrm{Unif}(0, x)$. Then $E(Y \mid X = x) = x/2$.

Example: breaking a stick

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

From the description of the experiment, $X \sim \mathrm{Unif}(0, 1)$ and $Y \mid X = x \sim \mathrm{Unif}(0, x)$. Then $E(Y \mid X = x) = x/2$.

By plugging in X for x, we have $E(Y \mid X) = X/2$.

Example: breaking a stick

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

From the description of the experiment, $X \sim \mathrm{Unif}(0, 1)$ and $Y \mid X = x \sim \mathrm{Unif}(0, x)$. Then $E(Y \mid X = x) = x/2$.

By plugging in X for x, we have $E(Y \mid X) = X/2$.

- The expected value of $E(Y \mid X)$ is

$$E(E(Y \mid X)) = E(X/2) = 1/4.$$

- The variance of $E(Y \mid X)$ is

$$\mathrm{Var}(E(Y \mid X)) = \mathrm{Var}(X/2) = 1/48.$$

Properties of conditional expectation

Conditional expectation has some very useful properties:

- If X and Y are *independent*, then $E(Y \mid X) = E(Y)$.
  - Independence implies $E(Y \mid X = x) = E(Y)$ for all x, hence $E(Y \mid X) = E(Y)$.
- Linearity:
  - $E(Y_1 + Y_2 \mid X) = E(Y_1 \mid X) + E(Y_2 \mid X)$
  - $E(cY \mid X) = cE(Y \mid X)$ for any constant c
- For any function h, $E(h(X)Y \mid X) = h(X)E(Y \mid X)$.
  - Conditional on X, functions of X act like a known constant.
- The **law of total expectation**: $E(E(Y \mid X)) = E(Y)$.

Conditional expectation under independence of RVs

If X and Y are independent, then $E(Y \mid X) = E(Y)$.

The converse is not always true.

Let $Z \sim \mathcal{N}(0,1)$ and $Y = Z^2$. Then $E(Y \mid Z) = E(Z^2 \mid Z) = Z^2$ and $E(Z \mid Y) = 0$.

- Conditional on $Y = y$, Z equals $\sqrt{y}$ or $-\sqrt{y}$ by the symmetry of the standard Normal, so $E(Z \mid Y = y) = 0$.

Despite the dependence between Z and Y, $E(Z \mid Y) = E(Z)$.

Law of total expectation

For any random variables X and Y,

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

Law of total expectation

For any random variables X and Y,

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

In the case where X and Y are both discrete, let $E(Y \mid X) = g(X)$. Applying LOTUS and expanding $g(x)$:

$$\begin{aligned}
E(g(X)) &= \sum_x g(x)P(X = x) \\
&= \sum_x \left(\sum_y yP(Y = y \mid X = x) \right) P(X = x) \\
&= \sum_x \sum_y yP(X = x)P(Y = y \mid X = x) \\
&= \sum_y y \sum_x P(X = x, Y = y) \\
&= \sum_y yP(Y = y) = E(Y).
\end{aligned}$$

Condition variance given a random variable

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

$$\mathrm{Var}(Y \mid X) = E((Y - E(Y|X))^2 \mid X).$$

This is equivalent to

$$\mathrm{Var}(Y \mid X) = E(Y^2 \mid X) - (E(Y \mid X))^2.$$

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

Example

Let $Z \sim \mathcal{N}(0,1)$ and $Y = Z^2$. Find $\mathrm{Var}(Y \mid Z)$ and $\mathrm{Var}(Z \mid Y)$.

Conditional on Z, Y is a known constant, so $\mathrm{Var}(Y \mid Z) = 0$. By the same reasoning, $\mathrm{Var}(h(Z) \mid Z) = 0$.

To get $\mathrm{Var}(Z \mid Y)$, apply the definition:

$$\mathrm{Var}(Z \mid Z^2) = E(Z^2 \mid Z^2) - (E(Z \mid Z^2))^2.$$

The first term equals Z^2. The second term equals 0 by symmetry. So $\mathrm{Var}(Z \mid Z^2) = Z^2$, or $\mathrm{Var}(Z \mid Y) = Y$.

Law of total variance

For any random variables X and Y,

$$\mathrm{Var}(Y) = E(\mathrm{Var}(Y \mid X)) + \mathrm{Var}(E(Y \mid X)).$$

Law of total variance

For any random variables X and Y,

$$\mathrm{Var}(Y) = E(\mathrm{Var}(Y \mid X)) + \mathrm{Var}(E(Y \mid X)).$$

Let $g(X) = E(Y \mid X)$. By the law of total expectation, $E(g(X)) = E(Y)$. Then

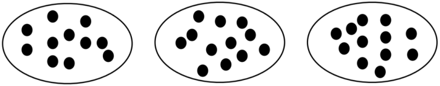
$$E(\mathrm{Var}(Y \mid X)) = E(E(Y^2 \mid X) - g(X)^2) = E(Y^2) - E(g(X)^2),$$

$$\mathrm{Var}(E(Y \mid X)) = E(g(X)^2) - (Eg(X))^2 = E(g(X)^2) - (EY)^2.$$

Adding these equations, we have the **law of total variance**. It is also known as the **variance decomposition formula**.

Law of total variance in terms of variance decomposition

Imagine a population where each person has a value of X (e.g., age) and a value of Y (e.g., height). We can divide this population into subpopulations, one for each possible value of X.



$X = 1$ $X = 2$ $X = 3$

There are two sources contributing to the variation in people's heights $\mathrm{Var}(Y)$ in the overall population.

- The **within-group variation** $E(\mathrm{Var}(Y \mid X))$: the average amount of variation in height within each age group.
- The **between-group variation** $\mathrm{Var}(E(Y \mid X))$: the variance of average heights across age groups.

The **law of total variance**: the total variance is the sum of within-group and between-group variation.

Law of total variance in terms of prediction

If we wanted to *predict* someone's height based on their age alone, the ideal scenario is *no* within-group variation in height.

- Within-group variation is also called **unexplained variation**.
- Between-group variation is also called **explained variation**.

The **law of total variance**: the overall variance of Y is the sum of unexplained and explained variation.

$$\begin{aligned}\mathrm{Var}(Y) &= E(\mathrm{Var}(Y \mid X)) + \mathrm{Var}(E(Y \mid X)) \\ &= \mathrm{Var}(Y - E(Y \mid X)) + \mathrm{Var}(E(Y \mid X)),\end{aligned}$$

since letting W be the residual $Y - E(Y \mid X)$,

$$\mathrm{Var}(Y - E(Y \mid X)) = E(W^2) = E(E(W^2 \mid X)) = E(\mathrm{Var}(Y \mid X)).$$

Random sum

A store receives N customers in a day, where N is a random variable with finite mean and variance. Let X_j be the amount spent by the jth customer at the store. Assume that each X_j has mean μ and variance σ^2, and that N and all the X_j's are independent of one another.

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

Random sum

A store receives N customers in a day, where N is a random variable with finite mean and variance. Let X_j be the amount spent by the jth customer at the store. Assume that each X_j has mean μ and variance σ^2, and that N and all the X_j's are independent of one another.

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

The random variable of interest is a **random sum**: the sum of a random number of random variables. There are two levels of randomness:

1. Each term in the sum is a random variable.
2. The number of terms in the sum is also a random variable.

Random sum, continued

Conditioning on N,

$$E(X \mid N) = E\left(\sum_{j=1}^{N} X_j \mid N\right) = \sum_{j=1}^{N} E(X_j \mid N) = \sum_{j=1}^{N} E(X_j) = N\mu,$$

$$\mathrm{Var}(X \mid N) = \mathrm{Var}\left(\sum_{j=1}^{N} X_j \mid N\right) = \sum_{j=1}^{N} \mathrm{Var}(X_j \mid N) = \sum_{j=1}^{N} \mathrm{Var}(X_j) = N\sigma^2.$$

Random sum, continued

Conditioning on N,

$$E(X \mid N) = E\left(\sum_{j=1}^{N} X_j \mid N\right) = \sum_{j=1}^{N} E(X_j \mid N) = \sum_{j=1}^{N} E(X_j) = N\mu,$$

$$\mathrm{Var}(X \mid N) = \mathrm{Var}\left(\sum_{j=1}^{N} X_j \mid N\right) = \sum_{j=1}^{N} \mathrm{Var}(X_j \mid N) = \sum_{j=1}^{N} \mathrm{Var}(X_j) = N\sigma^2.$$

By the law of total expectation,

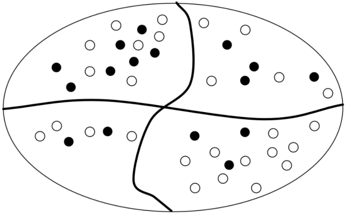
$$E(X) = E(E(X \mid N)) = E(N\mu) = \mu E(N).$$

By the law of total variance,

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

Random sample from a random city

To study the prevalence of a disease in several cities of interest within a certain county, we pick a city at random, then pick a random sample of n people from that city.



To illustrate, the oval-shaped county has 4 cities. Each city has healthy people (represented as white dots) and diseased people (black dots). A random city is chosen, and then a sample of n people are randomly selected from the chosen city for the study.

This is a form of a survey technique known as **cluster sampling**.

Random sample from a random city, continued

Let Q be the proportion of diseased people in the chosen city, and let X be the number of diseased people in the sample.

Suppose that $Q \sim \mathrm{Unif}(0, 1)$. Also assume that conditional on Q, each individual in the sample independently has probability Q of having the disease. Find $E(X)$ and $\mathrm{Var}(X)$.

Random sample from a random city, continued

Let Q be the proportion of diseased people in the chosen city, and let X be the number of diseased people in the sample.

Suppose that $Q \sim \mathrm{Unif}(0,1)$. Also assume that conditional on Q, each individual in the sample independently has probability Q of having the disease. Find $E(X)$ and $\mathrm{Var}(X)$.

There are two components to the variability in the number of diseased people in the sample:

1. Variation due to different cities having different disease prevalence.
2. Variation due to the randomness of the sample within the chosen city.

This is an example of a **multilevel model** (also known as **hierarchical model**).

Random sample from a random city, continued

Conditional on knowing the disease prevalence Q in the chosen city, each sampled individual is an independent Bernoulli trial with probability Q of success: $X \mid Q \sim \mathrm{Bin}(n, Q)$. So $E(X \mid Q) = nQ$ and $\mathrm{Var}(X \mid Q) = nQ(1 - Q)$.

Since Q is a standard Uniform random variable, $E(Q) = 1/2$, $E(Q^2) = 1/3$, and $\mathrm{Var}(Q) = 1/12$.

Random sample from a random city, continued

Conditional on knowing the disease prevalence Q in the chosen city, each sampled individual is an independent Bernoulli trial with probability Q of success: $X \mid Q \sim \mathrm{Bin}(n, Q)$. So $E(X \mid Q) = nQ$ and $\mathrm{Var}(X \mid Q) = nQ(1-Q)$.

Since Q is a standard Uniform random variable, $E(Q) = 1/2$, $E(Q^2) = 1/3$, and $\mathrm{Var}(Q) = 1/12$.

Applying the laws of total expectation and variance:

$$E(X) = E(E(X \mid Q)) = E(nQ) = \frac{n}{2},$$

$$\begin{aligned}\mathrm{Var}(X) &= E(\mathrm{Var}(X \mid Q)) + \mathrm{Var}(E(X \mid Q)) \\ &= E(nQ(1-Q)) + \mathrm{Var}(nQ) \\ &= nE(Q) - nE(Q^2) + n^2\mathrm{Var}(Q) \\ &= \frac{n}{6} + \frac{n^2}{12}.\end{aligned}$$