STAT 24400 Lecture 10
A Technique to Find Expectation & Variance
Section 4.4 Conditional Expectation & Prediction

Yibi Huang
Department of Statistics
University of Chicago

A Technique to find Expected Value & Variance

A Technique to find Expected Value & Variance

Sometimes it might be hard to find the exact distribution of a discrete random variable Y, but it's possible to express it as a sum of several random variables

$$Y = X_1 + X_2 + \cdots + X_n$$

that the distribution for X_i's are easier to find.

We can then find $\mathrm{E}(Y)$ and $\mathrm{Var}(Y)$ by

$$\mathrm{E}(Y) = \mathrm{E}(X_1) + \mathrm{E}(X_2) + \cdots + \mathrm{E}(X_n),$$

$$\mathrm{Var}(Y) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j),$$

even when the distribution of Y is unknown

An example is Coupon Collector's Problem on p.27-29 in L09.

Example (Random Hats Problem)

At a party, n men take off their hats.
The hats are then mixed up, and each man randomly grabs one.
Let Y be the number of men who grab their own hats.
Find $\mathrm{E}(Y)$ and $\mathrm{Var}(Y)$.

- Not trivial to find the PMF $\mathrm{P}(Y = k)$, $k = 0, 1, 2, \dots, n$.

Example (Random Hats Problem)

At a party, n men take off their hats.
The hats are then mixed up, and each man randomly grabs one.
Let Y be the number of men who grab their own hats.
Find $\mathrm{E}(Y)$ and $\mathrm{Var}(Y)$.

- Not trivial to find the PMF $\mathrm{P}(Y = k)$, $k = 0, 1, 2, \dots, n$.
- Nonetheless, we can find $\mathrm{E}(Y)$ and $\mathrm{Var}(Y)$ by writing Y as

$$Y = X_1 + X_2 + \cdots + X_n,$$

where

$$X_i = \begin{cases} 1, & \text{if the } i\text{th man grabs his own hat,} \\ 0, & \text{otherwise.} \end{cases}$$

- Note X_i's are **Bernoulli** but NOT independent

Expectation, Variance & Covariance of Bernoulli R.V.'s

For a Bernoulli Random Variable X with $p = \mathrm{P}(X = 1)$, it's expected value is

$$\mathrm{E}[X] = 1 \cdot \mathrm{P}(X = 1) + 0 \cdot \mathrm{P}(X = 0) = \mathrm{P}(X = 1) = p,$$

and the variance is

$$\mathrm{Var}(X_i) = p(1 - p) = \mathrm{P}(X = 1)\left(1 - \mathrm{P}(X = 1)\right).$$

Expectation, Variance & Covariance of Bernoulli R.V.'s

For a Bernoulli Random Variable X with $p = \mathrm{P}(X = 1)$, it's expected value is

$$\mathrm{E}[X] = 1 \cdot \mathrm{P}(X = 1) + 0 \cdot \mathrm{P}(X = 0) = \mathrm{P}(X = 1) = p,$$

and the variance is

$$\mathrm{Var}(X_i) = p(1 - p) = \mathrm{P}(X = 1)\,(1 - \mathrm{P}(X = 1))\,.$$

For two Bernoulli random variables X_i, X_j,

$$\begin{aligned}\mathrm{E}(X_i X_j) &= 1 \cdot 1 \cdot \mathrm{P}(X_i = 1, X_j = 1) + 1 \cdot 0 \cdot \mathrm{P}(X_i = 1, X_j = 0) \\ &\quad + 0 \cdot 1 \cdot \mathrm{P}(X_i = 0, X_j = 1) + 0 \cdot 0 \cdot \mathrm{P}(X_i = 0, X_j = 0) \\ &= \mathrm{P}(X_i = 1, X_j = 1),\end{aligned}$$

their covariance is thus

$$\begin{aligned}\mathrm{Cov}(X_i, X_j) &= \mathrm{E}(X_i X_j) - \mathrm{E}(X_i)\,\mathrm{E}(X_j) \\ &= \mathrm{P}(X_i = 1, X_j = 1) - \mathrm{P}(X_i = 1)\mathrm{P}(X_j = 1).\end{aligned}$$

Example (Random Hats Problem) — $\mathrm{E}(Y)$

As the ith man is equally likely to grab any of the n hats, it follows that

$$\mathrm{P}(X_i = 1) = \mathrm{P}(i\text{th man grabs his own hat}) = \frac{1}{n},$$

and so

$$\mathrm{E}[X_i] = \mathrm{P}(X_i = 1) = \frac{1}{n}.$$

Hence, we obtain

$$\mathrm{E}(Y) = \mathrm{E}(X_1) + \cdots + \mathrm{E}(X_n) = \underbrace{\frac{1}{n} + \cdots + \frac{1}{n}}_{n \text{ times}} = n \cdot \frac{1}{n} = 1.$$

Example (Random Hats Problem) — $\mathrm{E}(Y)$

As the ith man is equally likely to grab any of the n hats, it follows that

$$\mathrm{P}(X_i = 1) = \mathrm{P}(i\text{th man grabs his own hat}) = \frac{1}{n},$$

and so

$$\mathrm{E}[X_i] = \mathrm{P}(X_i = 1) = \frac{1}{n}.$$

Hence, we obtain

$$\mathrm{E}(Y) = \mathrm{E}(X_1) + \cdots + \mathrm{E}(X_n) = \underbrace{\frac{1}{n} + \cdots + \frac{1}{n}}_{n \text{ times}} = n \cdot \frac{1}{n} = 1.$$

We expect **only 1 man** can get his own hat if all men grab a hat randomly.

Example (Random Hats Problem) — $\mathrm{Cov}(X_i, X_j)$

$$\begin{aligned}
\mathrm{E}(X_iX_j) &= \mathrm{P}(X_i = 1, X_j = 1) \\
&= \mathrm{P}(X_i = 1)\mathrm{P}(X_j = 1 \mid X_i = 1) \\
&= \mathrm{P}(i\text{th man gets his hat}) \times \mathrm{P}(j\text{th man gets his hat} \mid i\text{th man gets his hat}) \\
&= \frac{1}{n} \cdot \frac{1}{n-1},
\end{aligned}$$

and thus their covariance is

$$\begin{aligned}
\mathrm{Cov}(X_i, X_j) &= \mathrm{E}(X_iX_j) - \mathrm{E}(X_i)\,\mathrm{E}(X_j) \\
&= \frac{1}{n(n-1)} - \frac{1}{n} \cdot \frac{1}{n} = \frac{1}{n^2(n-1)}.
\end{aligned}$$

Example (Random Hats Problem) — $\mathrm{Var}(Y)$

As X_i's are Bernoulli with $p = \frac{1}{n}$, their variance is

$$\mathrm{Var}(X_i) = p(1-p) = \frac{1}{n}\left(1 - \frac{1}{n}\right).$$

Example (Random Hats Problem) — $\mathrm{Var}(Y)$

As X_i's are Bernoulli with $p = \frac{1}{n}$, their variance is

$$\mathrm{Var}(X_i) = p(1-p) = \frac{1}{n}\left(1 - \frac{1}{n}\right).$$

Putting everything together, we get

$$\begin{aligned}
\mathrm{Var}(Y) &= \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j), \\
&= \sum_{i=1}^{n} \frac{1}{n}\left(1 - \frac{1}{n}\right) + 2\sum_{i<j} \frac{1}{n^2(n-1)} \\
&= n \cdot \frac{1}{n}\left(1 - \frac{1}{n}\right) + 2\binom{n}{2}\frac{1}{n^2(n-1)} \\
&= 1
\end{aligned}$$

Example (Random Hats Problem) — PMF (May Skip)

Just FYI, the PMF for $Y = \#$ of men who grab their own hats is

$$
\begin{aligned}
P(Y = n) &= \frac{1}{n!}, \\
P(Y = n-1) &= 0, \\
P(Y = 0) &= \sum_{i=2}^{n} \frac{(-1)^i}{i!} = \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \cdots + \frac{(-1)^n}{n!} \\
P(Y = k) &= \frac{1}{k!} \sum_{i=2}^{n-k} \frac{(-1)^i}{i!} = \frac{1}{k!} \left(\frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \cdots + \frac{(-1)^{n-k}}{(n-k)!} \right)
\end{aligned}
$$

for $k = 1, \ldots, n-2$.

See Example 5d on p.111-112 in
A First Course in Probability, 10ed, by Sheldon Ross
for calculation.

Example (Another Coupon Collector)

If each box of breakfast cereals contains a coupon,

- there are 25 different types of coupons,
- the coupon in any box is equally likely to be any of the 25 types,

Let $Y =$ the number of types of coupons in 10 boxes of cereals.
Find $\mathrm{E}(Y)$ and $\mathrm{Var}(Y)$.

Example (Another Coupon Collector)

If each box of breakfast cereals contains a coupon,

- there are 25 different types of coupons,
- the coupon in any box is equally likely to be any of the 25 types,

Let $Y =$ the number of types of coupons in 10 boxes of cereals.
Find $\mathrm{E}(Y)$ and $\mathrm{Var}(Y)$.

- Again, it's not trivial to find the PMF of Y

Example (Another Coupon Collector)

If each box of breakfast cereals contains a coupon,

- there are 25 different types of coupons,
- the coupon in any box is equally likely to be any of the 25 types,

Let Y = the number of types of coupons in 10 boxes of cereals.
Find $\mathrm{E}(Y)$ and $\mathrm{Var}(Y)$.

- Again, it's not trivial to find the PMF of Y
- Nonetheless, we can find $\mathrm{E}(Y)$ and $\mathrm{Var}(Y)$ by writing Y as

$$Y = X_1 + X_2 + \cdots + X_{25},$$

where

$$X_i = \begin{cases} 1, & \text{if at least one type } i \text{ coupon is in the 10 boxes,} \\ 0, & \text{otherwise.} \end{cases}$$

Example (Another Coupon Collector) — $\mathrm{E}(Y)$

$$\begin{aligned}\mathrm{E}[X_i] &= \mathrm{P}(X_i = 1)\\ &= \mathrm{P}(\text{at least one type } i \text{ coupon is in the 10 boxes})\\ &= 1 - \mathrm{P}(\text{no type } i \text{ coupons are in the 10 boxes})\\ &= 1 - \left(\frac{24}{25}\right)^{10}\end{aligned}$$

where the last equality follows since each of the 10 boxes will (independently) not contain a type i with probability $24/25$. Hence,

$$\mathrm{E}(Y) = \mathrm{E}(X_1) + \cdots + \mathrm{E}(X_{25}) = 25\left(1 - \left(\frac{24}{25}\right)^{10}\right) \approx 8.38.$$

Example (Another Coupon Collector) — $\mathrm{Cov}(X_i, X_j)$

It's easier to find

$$\mathrm{P}(X_i = 0, X_j = 0) = \mathrm{P}(\text{No Type } i \text{ or } j \text{ coupons in 10 boxes}) = \left(\frac{23}{25}\right)^{10},$$

than to find

$$\mathrm{P}(X_i = 1, X_j = 1) = \mathrm{P}(\text{Both Types } i \text{ and } j \text{ are in 10 boxes}).$$

Example (Another Coupon Collector) — $\mathrm{Cov}(X_i, X_j)$

It's easier to find

$$\mathrm{P}(X_i = 0, X_j = 0) = \mathrm{P}(\text{No Type } i \text{ or } j \text{ coupons in 10 boxes}) = \left(\frac{23}{25}\right)^{10},$$

than to find

$$\mathrm{P}(X_i = 1, X_j = 1) = \mathrm{P}(\text{Both Types } i \text{ and } j \text{ are in 10 boxes}).$$

Can we find $\mathrm{Cov}(X_i, X_j)$ using $\mathrm{P}(X_i = 0, X_j = 0)$?

Example (Another Coupon Collector) — $\mathrm{Cov}(X_i, X_j)$

It's easier to find

$$\mathrm{P}(X_i = 0, X_j = 0) = \mathrm{P}(\text{No Type } i \text{ or } j \text{ coupons in 10 boxes}) = \left(\frac{23}{25}\right)^{10},$$

than to find

$$\mathrm{P}(X_i = 1, X_j = 1) = \mathrm{P}(\text{Both Types } i \text{ and } j \text{ are in 10 boxes}).$$

Can we find $\mathrm{Cov}(X_i, X_j)$ using $\mathrm{P}(X_i = 0, X_j = 0)$?

Yes. Let $Z_i = 1 - X_i$, then $\mathrm{Cov}(Z_i, Z_j) = \mathrm{Cov}(1 - X_i, 1 - X_j) = \mathrm{Cov}(X_i, X_j)$, and

$$\begin{aligned}
\mathrm{Cov}(Z_i, Z_j) &= \mathrm{E}(Z_i Z_j) - \mathrm{E}(Z_i)\,\mathrm{E}(Z_j) \\
&= \mathrm{P}(Z_i = 1, Z_j = 1) - \mathrm{P}(Z_i = 1)\mathrm{P}(Z_j = 1) \\
&= \mathrm{P}(X_i = 0, X_j = 0) - \mathrm{P}(X_i = 0)\mathrm{P}(X_j = 0) \\
&= \left(\frac{23}{25}\right)^{10} - \left(\frac{24}{25}\right)^{20} \approx -0.007614.
\end{aligned}$$

Example (Another Coupon Collector) — $\mathrm{Var}(Y)$

As X_i's are Bernoulli with $p = 1 - (24/25)^{10}$, their variance is

$$\mathrm{Var}(X_i) = p(1-p) = \left(\frac{24}{25}\right)^{10} \left(1 - \left(\frac{24}{25}\right)^{10}\right) \approx 0.22283.$$

Example (Another Coupon Collector) — $\mathrm{Var}(Y)$

As X_i's are Bernoulli with $p = 1 - (24/25)^{10}$, their variance is

$$\mathrm{Var}(X_i) = p(1-p) = \left(\frac{24}{25}\right)^{10}\left(1 - \left(\frac{24}{25}\right)^{10}\right) \approx 0.22283.$$

Putting everything together, we get

$$\begin{aligned}\mathrm{Var}(Y) &= \sum_{i=1}^{25} \mathrm{Var}(X_i) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j), \\ &\approx 25 \times 0.22283 + 2\binom{25}{2}(-0.007614) \\ &\approx 1.0024.\end{aligned}$$

Example (Coin Flip Pattern HTTH) — $\mathrm{E}(Y)$

Let Y be the total number of times that you see the pattern HTTH in n flips of a fair coin. Find $\mathrm{E}(Y)$ and $\mathrm{Var}(Y)$.

Note the coin flip sequences with the pattern can overlap.
e.g., the pattern HTTH shows up twice, not once, in the sequence

$$\mathtt{H}\,\overbrace{\mathtt{H\,T\,T\,H}}^{\text{first}}\!\!\!\!\!\!\underbrace{\mathtt{H\,T\,T\,H}}_{\text{second}}\,\mathtt{H\,T}$$

Sol. Let $(C_1, \dots, C_n)$ be the outcome of the n flips. Writing Y as

$$Y = X_1 + X_2 + \cdots + X_{n-3}, \quad \text{where } X_i = \begin{cases} 1, & \text{if } (C_i, C_{i+1}, C_{i+2}, C_{i+3}) = \mathtt{HTTH}, \\ 0, & \text{otherwise.} \end{cases}$$

As $\mathrm{E}[X_i] = \mathrm{P}(X_i = 1) = \mathrm{P}((C_i, C_{i+1}, C_{i+2}, C_{i+3}) = \mathtt{HTTH}) = (1/2)^4$,
we get $\mathrm{E}(Y) = \mathrm{E}(X_1) + \cdots + \mathrm{E}(X_{n-3}) = (n-3)(1/2)^4$.

Example (Coin Flip Pattern HTTH) — $\mathrm{Cov}(X_i, X_j)$

- $\mathrm{Cov}(X_i, X_j) = 0$ if $|i - j| > 3$ as $(C_i, C_{i+1}, C_{i+2}, C_{i+3})$ and $(C_j, C_{j+1}, C_{j+2}, C_{j+3})$ are independent if $|i - j| > 3$.

Example (Coin Flip Pattern HTTH) — $\mathrm{Cov}(X_i, X_j)$

- $\mathrm{Cov}(X_i, X_j) = 0$ if $|i - j| > 3$ as $(C_i, C_{i+1}, C_{i+2}, C_{i+3})$ and $(C_j, C_{j+1}, C_{j+2}, C_{j+3})$ are independent if $|i - j| > 3$.
- $\mathrm{E}(X_i X_{i+1}) = \mathrm{P}(X_i = 1, X_{i+1} = 1) = 0$ since if $(C_i, C_{i+1}, C_{i+2}, C_{i+3}) =$ HTTH, then $(C_{i+1}, C_{i+2}, C_{i+3}, C_{i+4})$ would be TTH?, not HTTH.

Example (Coin Flip Pattern HTTH) — $\mathrm{Cov}(X_i, X_j)$

- $\mathrm{Cov}(X_i, X_j) = 0$ if $|i - j| > 3$ as $(C_i, C_{i+1}, C_{i+2}, C_{i+3})$ and $(C_j, C_{j+1}, C_{j+2}, C_{j+3})$ are independent if $|i - j| > 3$.
- $\mathrm{E}(X_i X_{i+1}) = \mathrm{P}(X_i = 1, X_{i+1} = 1) = 0$ since if $(C_i, C_{i+1}, C_{i+2}, C_{i+3}) =$ HTTH, then $(C_{i+1}, C_{i+2}, C_{i+3}, C_{i+4})$ would be TTH?, not HTTH. It follows that

$$\mathrm{Cov}(X_i, X_{i+1}) = \mathrm{E}(X_i X_{i+1}) - \mathrm{E}(X_i)\,\mathrm{E}(X_{i+1}) = 0 - (1/2)^8 = \frac{-1}{256}$$

Example (Coin Flip Pattern HTTH) — $\mathrm{Cov}(X_i, X_j)$

- $\mathrm{Cov}(X_i, X_j) = 0$ if $|i - j| > 3$ as $(C_i, C_{i+1}, C_{i+2}, C_{i+3})$ and $(C_j, C_{j+1}, C_{j+2}, C_{j+3})$ are independent if $|i - j| > 3$.
- $\mathrm{E}(X_i X_{i+1}) = \mathrm{P}(X_i = 1, X_{i+1} = 1) = 0$ since if $(C_i, C_{i+1}, C_{i+2}, C_{i+3}) = \mathtt{HTTH}$, then $(C_{i+1}, C_{i+2}, C_{i+3}, C_{i+4})$ would be TTH?, not HTTH. It follows that

$$\mathrm{Cov}(X_i, X_{i+1}) = \mathrm{E}(X_i X_{i+1}) - \mathrm{E}(X_i)\,\mathrm{E}(X_{i+1}) = 0 - (1/2)^8 = \frac{-1}{256}$$

- Likewise, $\mathrm{Cov}(X_i, X_{i+2}) = \dfrac{-1}{256}$.

Example (Coin Flip Pattern HTTH) — $\mathrm{Cov}(X_i, X_j)$

- $\mathrm{Cov}(X_i, X_j) = 0$ if $|i-j| > 3$ as $(C_i, C_{i+1}, C_{i+2}, C_{i+3})$ and $(C_j, C_{j+1}, C_{j+2}, C_{j+3})$ are independent if $|i-j| > 3$.
- $\mathrm{E}(X_i X_{i+1}) = \mathrm{P}(X_i = 1, X_{i+1} = 1) = 0$ since if $(C_i, C_{i+1}, C_{i+2}, C_{i+3}) = \mathtt{HTTH}$, then $(C_{i+1}, C_{i+2}, C_{i+3}, C_{i+4})$ would be TTH?, not HTTH. It follows that

$$\mathrm{Cov}(X_i, X_{i+1}) = \mathrm{E}(X_i X_{i+1}) - \mathrm{E}(X_i)\,\mathrm{E}(X_{i+1}) = 0 - (1/2)^8 = \frac{-1}{256}$$

- Likewise, $\mathrm{Cov}(X_i, X_{i+2}) = \dfrac{-1}{256}$.
- $\mathrm{E}(X_i X_{i+3}) = \mathrm{P}(X_i = 1, X_{i+3} = 1) = (1/2)^7$ since $(C_i, C_{i+1}, C_{i+2}, C_{i+3}) = \mathtt{HTTH}$ and $(C_{i+3}, C_{i+4}, C_{i+5}, C_{i+6}) = \mathtt{HTTH}$ implies

$$(C_i, C_{i+1}, C_{i+2}, C_{i+3}, C_{i+4}, C_{i+5}, C_{i+6}) = \mathtt{HTTHTTH},$$

and thus

$$\mathrm{Cov}(X_i, X_{i+3}) = \mathrm{E}(X_i X_{i+3}) - \mathrm{E}(X_i)\,\mathrm{E}(X_{i+3}) = (1/2)^7 - (1/2)^8 = \frac{1}{256}.$$

Example (Coin Flip Pattern HTTH) — $\mathrm{Var}(Y)$

As X_i's are Bernoulli with $p = (1/2)^4 = 1/16$, their variance is

$$\mathrm{Var}(X_i) = p(1-p) = \frac{15}{256}.$$

Example (Coin Flip Pattern HTTH) — $\mathrm{Var}(Y)$

As X_i's are Bernoulli with $p = (1/2)^4 = 1/16$, their variance is

$$\mathrm{Var}(X_i) = p(1-p) = \frac{15}{256}.$$

Putting everything together, we get

$$\begin{aligned}
\mathrm{Var}(Y) &= \sum_{i=1}^{n-3} \mathrm{Var}(X_i) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j), \\
&= (n-3)\,\mathrm{Var}(X_i) + 2(n-4)\,\mathrm{Cov}(X_i, X_{i+1}) \\
&\quad + 2(n-5)\,\mathrm{Cov}(X_i, X_{i+2}) + 2(n-6)\,\mathrm{Cov}(X_i, X_{i+3}) \\
&= (n-3)\frac{15}{256} + 2(n-4)(\frac{-1}{256}) + 2(n-5)(\frac{-1}{256}) + 2(n-6)(\frac{1}{256}) \\
&= (n-3)\frac{13}{256}
\end{aligned}$$

Conditional Expectation and Prediction

Conditional Expectation

For two random variables X and Y, the *conditional mean* or *conditional expected value of* Y given $X = x$ is defined to be

$$\mu_{Y|X=X} = \mathrm{E}(Y \mid X = x) = \begin{cases} \sum_y y\, p_{Y|X}(y \mid x) & \text{if discrete} \\ \int_{-\infty}^{\infty} y\, f_{Y|X}(y \mid x) dy & \text{if continuous} \end{cases}$$

where $p_{Y|X}(y \mid x)$ and $f_{Y|X}(y \mid x)$ are the conditional PMF/PDF of Y given X.

Conditional Expectation

For two random variables X and Y, the *conditional mean* or *conditional expected value of* Y given $X = x$ is defined to be

$$\mu_{Y|X=X} = \mathrm{E}(Y \mid X = x) = \begin{cases} \sum_y y\, p_{Y|X}(y \mid x) & \text{if discrete} \\ \int_{-\infty}^{\infty} y\, f_{Y|X}(y \mid x) dy & \text{if continuous} \end{cases}$$

where $p_{Y|X}(y \mid x)$ and $f_{Y|X}(y \mid x)$ are the conditional PMF/PDF of Y given X.

- The conditional mean $\mathrm{E}(Y \mid X = x)$ is NOT a single value but a function of the x value given.
  $\Rightarrow$ $\mathrm{E}(Y \mid X)$ is a function $h(X)$ of X and thus is a random variable.

Conditional Expectation

For two random variables X and Y, the *conditional mean* or *conditional expected value of* Y given $X = x$ is defined to be

$$\mu_{Y|X=X} = \mathrm{E}(Y \mid X = x) = \begin{cases} \sum_y y\, p_{Y|X}(y \mid x) & \text{if discrete} \\ \int_{-\infty}^{\infty} y\, f_{Y|X}(y \mid x) dy & \text{if continuous} \end{cases}$$

where $p_{Y|X}(y \mid x)$ and $f_{Y|X}(y \mid x)$ are the conditional PMF/PDF of Y given X.

- The conditional mean $\mathrm{E}(Y \mid X = x)$ is NOT a single value but a function of the x value given.
  $\Rightarrow$ $\mathrm{E}(Y \mid X)$ is a function $h(X)$ of X and thus is a random variable.

More generally, the *conditional mean* or *conditional expected value of* $g(Y)$ *given* $X = x$ is

$$\mathrm{E}(g(Y) \mid X = x) = \begin{cases} \sum_y g(y)\, p_{Y|X}(y \mid x) & \text{if discrete} \\ \int_{-\infty}^{\infty} g(y)\, f_{Y|X}(y \mid x) dy & \text{if continuous} \end{cases}$$

Conditional Variance

We can also define the *conditional variance* of Y given $X = x$.

$$\mathrm{Var}(Y \mid X = x) = \mathrm{E}\left([Y - \mathrm{E}(Y \mid X = x)]^2 \mid X = x\right)$$

Shortcut formula for conditional variance:

$$\mathrm{Var}(Y \mid X = x) = \mathrm{E}(Y^2 \mid X = x) - \left[\mathrm{E}(Y \mid X = x)\right]^2$$

Example (Gas Station) — Conditional Mean

Recall in L06, the conditional PMF of Y given $X = x$ is as follows.

conditional PMF	$p(y \mid x)$	Y = 0	1	2
X	0	0.625	0.25	0.125
	1	0.2353	0.5882	0.1765
	2	0.12	0.28	0.60

The conditional mean of Y given $X = x$ is

$$\mathrm{E}(Y \mid X = x) = \begin{cases} 0 \cdot 0.625 + 1 \cdot 0.25 + 2 \cdot 0.125 = 0.5 & \text{if } x = 0 \\ 0 \cdot 0.2353 + 1 \cdot 0.5882 + 2 \cdot 0.1765 = 0.9412 & \text{if } x = 1 \\ 0 \cdot 0.12 + 1 \cdot 0.28 + 2 \cdot 0.6 = 1.48 & \text{if } x = 2 \end{cases}$$

Example — Poisson

For independent r.v.'s $X_1 \sim \text{Poisson}(\lambda_1)$ and $X_2 \sim \text{Poisson}(\lambda_2)$, we showed on p.18 in L06 that, given $T = X_1 + X_2 = t$, the conditional distribution of X_1 is

$$X_1 \,|_{T=t} \sim \text{Bin}\left(t, \frac{\lambda_1}{\lambda_1 + \lambda_2}\right).$$

As the expected value for $\text{Bin}(n, p)$ is np, and the variance is $np(1-p)$, it follows that

$$\mathrm{E}(X_1 \mid T) = \frac{\lambda_1}{\lambda_1 + \lambda_2} T \quad \mathrm{Var}(X_1 \mid T) = \frac{\lambda_1 \lambda_2}{(\lambda_1 + \lambda_2)^2} T$$

- Note that $\mathrm{E}(X_1 \mid T) = \frac{\lambda_1}{\lambda_1+\lambda_2} T$ and $\mathrm{Var}(X_1 \mid T) = \frac{\lambda_1\lambda_2}{(\lambda_1+\lambda_2)^2} T$ are both functions of T and hence random variables.

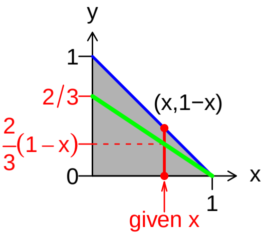
Example (Mixed Nuts) — Conditional Mean

Recall in L06, we found the conditional PDF $f_{Y|X}(y \mid x)$ of Y (cashew) given $X = x$ (almond) to be

$$f_{Y|X}(y \mid x) = \frac{f(x,y)}{f_X(x)} = \frac{2y}{(1-x)^2}, \quad \text{for } 0 \le y \le 1 - x\,.$$

The conditional expected weight of Y (cashew) in a can given there being $X = x$ lbs of almond in the can is

$$\begin{aligned} \mathrm{E}(Y \mid X = x) &= \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x) dy \\ &= \int_0^{1-x} y \cdot \frac{2y}{(1-x)^2} dy \\ &= \frac{2y^3}{3(1-x)^2}\Bigg|_{y=0}^{y=1-x} = \frac{2}{3}(1-x). \end{aligned}$$



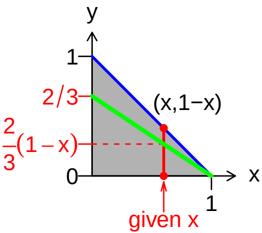
Example (Mixed Nuts) — Conditional Mean

Recall in L06, we found the conditional PDF $f_{Y|X}(y \mid x)$ of Y (cashew) given $X = x$ (almond) to be

$$f_{Y|X}(y \mid x) = \frac{f(x,y)}{f_X(x)} = \frac{2y}{(1-x)^2}, \quad \text{for } 0 \le y \le 1-x\,.$$

The conditional expected weight of Y (cashew) in a can given there being $X = x$ lbs of almond in the can is

$$\begin{aligned} \mathrm{E}(Y \mid X = x) &= \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x) dy \\ &= \int_0^{1-x} y \cdot \frac{2y}{(1-x)^2} dy \\ &= \frac{2y^3}{3(1-x)^2}\Bigg|_{y=0}^{y=1-x} = \frac{2}{3}(1-x). \end{aligned}$$



- Note that $\mathrm{E}(Y \mid X) = \frac{2}{3}(1-X)$ is a function of X and is thus a random variable.

Example (Mixed Nuts) — Conditional Variance

$$\mathrm{E}(Y^2 \mid X = x) = \int_{-\infty}^{\infty} y^2 f_{Y|X}(y \mid x) dy = \int_0^{1-x} y^2 \cdot \frac{2y}{(1-x)^2} dy$$

$$= \frac{y^4}{2(1-x)^2}\Bigg|_{y=0}^{y=1-x} = \frac{1}{2}(1-x)^2.$$

So

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

$$= \frac{1}{2}(1-x)^2 - \left(\frac{2}{3}(1-x)\right)^2 = \frac{1}{18}(1-x)^2$$

- Note $\mathrm{Var}(Y \mid X) = \frac{1}{18}(1-X)^2$ is a a function of X and is thus a random variable.

Tower Law $\mathrm{E}(\mathrm{E}(Y \mid X)) = \mathrm{E}(Y)$

As $\mathrm{E}(Y \mid X)$ is a random variable and it's a function of X we can take its expected value and it can be shown that

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

This is called the *Tower Law*, or *Law of Total Expectation*.

$$\underbrace{\mathrm{E}_X(\quad \overbrace{\mathrm{E}_{Y|X}(Y \mid X)}^{\text{taking expectation over condi. distn. of } Y \text{ given } X} \quad)}_{\text{taking expectation over marginal distn. of } X} = \mathrm{E}(Y).$$

- Tower Law is useful when it's hard to find the marginal distribution of Y, but easy to find $\mathrm{E}_{Y|X}(Y \mid X)$.

Example (Gas Station) — Tower Law

The conditional mean of Y given $X = x$ is

x	0	1	2
$\mathrm{E}(Y \mid X = x)$	0.5	0.9412	1.48

and the marginal PMFs for X was obtained in L05 to be

x	0	1	2
$p_X(x)$	0.16	0.34	0.50

It follows that

$$\begin{aligned} \mathrm{E}_X(\mathrm{E}(Y \mid X)) &= \mathrm{E}(Y \mid X = 0)p_X(0) + \mathrm{E}(Y \mid X = 1)p_X(1) + \mathrm{E}(Y \mid X = 2)p_X(2) \\ &= 0.5 \cdot 0.16 + 0.9412 \cdot 0.34 + 1.48 \cdot 0.50 = 1.14. \end{aligned}$$

which is identical to $\mathrm{E}(Y)$ computed using the marginal PMF of Y

y	0	1	2
$p_Y(y)$	0.24	0.38	0.38

$$\mathrm{E}(Y) = 0 \cdot 0.24 + 1 \cdot 0.38 + 2 \cdot 0.38 = 1.14.$$

Example — Poisson

For independent r.v.'s $X_1 \sim \text{Poisson}(\lambda_1)$ and $X_2 \sim \text{Poisson}(\lambda_2)$, and $T = X_1 + X_2$, recall we showed earlier the conditional mean of X_1 given T is

$$\mathrm{E}(X_1 \mid T) = \frac{\lambda_1 T}{\lambda_1 + \lambda_2}.$$

Example — Poisson

For independent r.v.'s $X_1 \sim \text{Poisson}(\lambda_1)$ and $X_2 \sim \text{Poisson}(\lambda_2)$, and $T = X_1 + X_2$, recall we showed earlier the conditional mean of X_1 given T is

$$\mathrm{E}(X_1 \mid T) = \frac{\lambda_1 T}{\lambda_1 + \lambda_2}.$$

Also recall that $T = X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$,and thus $\mathrm{E}(T) = \lambda_1 + \lambda_2$.
It follows that

$$\mathrm{E}[\mathrm{E}(X_1 \mid T)] = \mathrm{E}\left[\frac{\lambda_1 T}{\lambda_1 + \lambda_2}\right] = \frac{\lambda_1 \mathrm{E}[T]}{\lambda_1 + \lambda_2} = \frac{\lambda_1(\lambda_1 + \lambda_2)}{\lambda_1 + \lambda_2} = \lambda_1 = \mathrm{E}(X_1).$$

Tower Law is also valid for this example.

Proof of Tower Law (Discrete Case)

$$
\begin{aligned}
\mathrm{E}(\mathrm{E}(Y \mid X)) &= \sum_x \underbrace{\mathrm{E}(Y \mid X = x)}_{\searrow} p_X(x) \\
&= \sum_x \overbrace{\sum_y y \cdot p_{Y|X}(y \mid x)} p_X(x) \\
&= \sum_x \sum_y y \cdot \frac{p_{XY}(x,y)}{p_X(x)} p_X(x) \\
&= \sum_x \sum_y y \cdot p_{XY}(x,y) \\
&= \sum_y y \underbrace{\sum_x p_{XY}(x,y)}_{p_Y(y)} \\
&= \sum_y y \cdot p_Y(y) = \mathrm{E}(Y)
\end{aligned}
$$

Proof for the continuous case is similar.

Sum of a Random Number of Random Variables

Consider sum of the type

$$T = \sum_{i=1}^{N} X_i,$$

where

- $X_1, X_2, \ldots$ are i.i.d. with $\mathrm{E}\,|X_i| < \infty$, and
- N is a non-negative integer-valued random variable, independent of X_i's.
- If $N = 0$, the sum is 0.

Ex: Let N be the number of claims an insurance company receives in a given month, and the amounts of the individual claims $X_1, X_2, \ldots$ are i.i.d. The total amount of claims in the month is then $\sum_{i=1}^{N} X_i$.

Expected Sum of a Random Number of Random Variables

If $N = n$ is a constant, we know

$$\mathrm{E}[T] = \mathrm{E}\left[\sum\nolimits_{i=1}^{n} X_i\right] = n\,\mathrm{E}(X),$$

where $\mathrm{E}(X)$ is the common mean of X_i's.

For random N, we can first find the conditional expected sum given $N = n$,

$$\mathrm{E}[T \mid N = n] = \mathrm{E}\left[\sum\nolimits_{i=1}^{N} X_i \,\middle|\, N = n\right] = \sum_{i=1}^{n} \underbrace{\mathrm{E}\left[X_i \mid N = n\right]}_{\substack{= \mathrm{E}[X] \text{ by indep.} \\ \text{of } N \text{ and } X_i\text{'s}}} = n\,\mathrm{E}[X]$$

i.e., $\mathrm{E}[T \mid N] = N\,\mathrm{E}[X]$. Applying the Tower Law, we get

$$\mathrm{E}[T] = \mathrm{E}\Bigg[\underbrace{\mathrm{E}[T \mid N]}_{= N\,\mathrm{E}[X]}\Bigg] = \mathrm{E}[N\,\underbrace{\mathrm{E}[X]}_{\text{constant}}] = \mathrm{E}[N]\,\mathrm{E}[X].$$

Example (a Game)

Find the expected reward for the following game: at each round, you toss a coin.

- If it's Heads, you roll a die and win $1 if you rolled a 6.
- If it's Tails, the game ends

Sol. Your total reward is $T = \sum_{i=1}^{N} X_i$, where

- X_i's are i.i.d. Bernoulli$(1/6)$, $\Rightarrow \mathrm{E}[X_i] = 1/6$.
- $N = \#$ of consecutive `H`'s obtained before getting the first `T`
  - Observe that $M = N + 1$ is Geometric$(p = 1/2)$,
    $\Rightarrow\ \mathrm{E}[M] = 1/p = 2\ \Rightarrow\ \mathrm{E}[N] = 1$.
- So your expected total reward is

$$\mathrm{E}[T] = \mathrm{E}[N]\,\mathrm{E}[X] = 1 \times (1/6) = 1/6.$$

Example (Mouse Trapped in a Maze)

A mouse is placed at a room in a maze containing 3 doors.

- Door #1 leads to a path that will lead it to freedom after 6 minutes of travel.
- Door #2 leads to a path that will return it to the same after 4 minutes of travel.
- Door #3 leads to a path that will return it to the same room after 2 minutes of travel.

Suppose the mouse always randomly chooses one of the 3 doors equally likely whenever it returns to the room it started. What is the expected length of time it takes the mouse to get free?

Example (Mouse Trapped in a Maze)

The total amount of time before the mouse gets free can be written as

$$T = \sum_{i=1}^{N} X_i$$

where

$$X_i = \text{travel time (in minutes) for its } i\text{th departure}$$
$$N = \# \text{ of departures from the room until free}$$

- Observe N is Geometric$(p = 1/3)$, so $\mathrm{E}[N] = 1/p = 3$.
- Clearly, X_i's are i.i.d. with the PMF

x	2	4	6
$p(x)$	1/3	1/3	1/3

$$\Rightarrow \mathrm{E}[X] = \frac{2+4+6}{3} = 4.$$

- However, X_i's and N are NOT independent.
  - $X_N = 6$, always
  - For $i < N$, X_i is equally likely to be 4 or 2, implying

$$\mathrm{E}[X_i \mid N = n] = \frac{4+2}{2} = 3, \quad i < n.$$

Example (Mouse Trapped in a Maze)

So

$$\begin{aligned}
\mathrm{E}[T \mid N = n] &= \mathrm{E}\left[\sum\nolimits_{i=1}^{N} X_i \,\middle|\, N = n\right] \\
&= \sum\nolimits_{i=1}^{n-1} \underbrace{\mathrm{E}[X_i \mid N = n]}_{=3} + \underbrace{\mathrm{E}[X_n \mid N = n]}_{=6} \\
&= 3(n-1) + 6 = 3n + 3.
\end{aligned}$$

As $\mathrm{E}[T \mid N] = 3N + 3$, apply the Tower Law and recall $\mathrm{E}[N] = 3$, we have

$$\mathrm{E}[T] = \mathrm{E}[\mathrm{E}[T \mid N]] = \mathrm{E}[3N + 3] = 3\,\mathrm{E}[N] + 3 = 3 \times 3 + 3 = 12.$$

On average, it takes the mouse 12 minutes to escape.

Example (Mouse Trapped in a Maze)

So

$$\begin{aligned} \mathrm{E}[T \mid N = n] &= \mathrm{E}\left[\sum_{i=1}^{N} X_i \,\middle|\, N = n\right] \\ &= \sum_{i=1}^{n-1} \underbrace{\mathrm{E}[X_i \mid N = n]}_{=3} + \underbrace{\mathrm{E}[X_n \mid N = n]}_{=6} \\ &= 3(n-1) + 6 = 3n + 3. \end{aligned}$$

As $\mathrm{E}[T \mid N] = 3N + 3$, apply the Tower Law and recall $\mathrm{E}[N] = 3$, we have

$$\mathrm{E}[T] = \mathrm{E}[\mathrm{E}[T \mid N]] = \mathrm{E}[3N + 3] = 3\,\mathrm{E}[N] + 3 = 3 \times 3 + 3 = 12.$$

On average, it takes the mouse 12 minutes to escape.

Remark: Note in this example,

$$\underbrace{\mathrm{E}\left[\sum_{i=1}^{N} X_i \,\middle|\, N = n\right]}_{=3n+3} \neq \underbrace{\mathrm{E}\left[\sum_{i=1}^{n} X_i\right]}_{=n\,\mathrm{E}[X_i]=4n}$$

Tower Law for Functions of X, Y

Tower Law not only works for Y itself, but also

- for any function $g(Y)$ of Y:

$$\mathrm{E}_X\left[\mathrm{E}_{Y|X}(g(Y) \mid X)\right] = \mathrm{E}(g(Y)),$$

- as well as for any function $h(X, Y)$ of X, Y:

$$\mathrm{E}_X[\mathrm{E}_{Y|X}(h(X, Y) \mid X)] = \mathrm{E}(h(X, Y)).$$

For example, $\mathrm{E}(\mathrm{E}(Y^2 \mid X)) = \mathrm{E}(Y^2)$.

$\mathrm{E}(\mathrm{Var}(Y \mid X))$ & $\mathrm{Var}(\mathrm{E}(Y \mid X))$

- $\mathrm{E}(Y \mid X)$ is a random variable,
  - its expected value is $\mathrm{E}[\mathrm{E}(Y \mid X)] = \mathrm{E}[Y]$ by Tower Law
  - its variance is

$$\underbrace{\mathrm{Var}_X\left(\overbrace{\mathrm{E}_{Y|X}(Y \mid X)}^{\text{taking expectation over condi. distn. of } Y \text{ given } X}\right)}_{\text{taking variance over marginal distn. of } X}$$

$\mathrm{E}(\mathrm{Var}(Y \mid X))$ & $\mathrm{Var}(\mathrm{E}(Y \mid X))$

- $\mathrm{E}(Y \mid X)$ is a random variable,
  - its expected value is $\mathrm{E}[\mathrm{E}(Y \mid X)] = \mathrm{E}[Y]$ by Tower Law
  - its variance is

$$\underbrace{\mathrm{Var}_X\left(\overbrace{\mathrm{E}_{Y|X}(Y \mid X)}^{\text{taking expectation over condi. distn. of } Y \text{ given } X}\right)}_{\text{taking variance over marginal distn. of } X}$$

- As $\mathrm{Var}(Y \mid X)$ is also a random variable, we can take expected value of it

$$\underbrace{\mathrm{E}_X\left(\overbrace{\mathrm{Var}_{Y|X}(Y \mid X)}^{\text{taking variance over condi. distn. of } Y \text{ given } X}\right)}_{\text{taking expectation over marginal distn. of } X}$$

Tower Law for Variance = Law of Total Variance

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

Intuitively, if Y has high variance, it comes from one of two sources:

- Either Y is highly variable even if you already know the value of X
- Or if not, then expected value of Y must change a lot as you var

Example: Poisson — Tower Law for Variance

For independent r.v.'s $X_1 \sim \text{Poisson}(\lambda_1)$, $X_2 \sim \text{Poisson}(\lambda_2)$, and $T = X_1 + X_2$, recall we showed earlier that

$$\mathrm{E}(X_1 \mid T) = \frac{\lambda_1}{\lambda_1 + \lambda_2} T, \quad \mathrm{Var}(X_1 \mid T) = \frac{\lambda_1 \lambda_2}{(\lambda_1 + \lambda_2)^2} T$$

Example: Poisson — Tower Law for Variance

For independent r.v.'s $X_1 \sim \text{Poisson}(\lambda_1)$, $X_2 \sim \text{Poisson}(\lambda_2)$, and $T = X_1 + X_2$, recall we showed earlier that

$$\mathrm{E}(X_1 \mid T) = \frac{\lambda_1}{\lambda_1 + \lambda_2} T, \quad \mathrm{Var}(X_1 \mid T) = \frac{\lambda_1 \lambda_2}{(\lambda_1 + \lambda_2)^2} T$$

Recall $T \sim \text{Poisson}(\lambda_1 + \lambda_2)$ which implies $\mathrm{E}[T] = \lambda_1 + \lambda_2$ and $\mathrm{Var}(T) = \lambda_1 + \lambda_2$. It follows that

$$\mathrm{Var}(\mathrm{E}(X_1 \mid T)) = \mathrm{Var}\left(\frac{\lambda_1}{\lambda_1 + \lambda_2} T\right) = \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^2 \underbrace{\mathrm{Var}(T)}_{=\lambda_1 + \lambda_2} = \frac{\lambda_1^2}{\lambda_1 + \lambda_2}$$

$$\mathrm{E}[\mathrm{Var}(X_1 \mid T)] = \mathrm{E}\left[\frac{\lambda_1 \lambda_2}{(\lambda_1 + \lambda_2)^2} T\right] = \frac{\lambda_1 \lambda_2}{(\lambda_1 + \lambda_2)^2} \underbrace{\mathrm{E}[T]}_{=\lambda_1 + \lambda_2} = \frac{\lambda_1 \lambda_2}{\lambda_1 + \lambda_2}$$

Example: Poisson — Tower Law for Variance

For independent r.v.'s $X_1 \sim \text{Poisson}(\lambda_1)$, $X_2 \sim \text{Poisson}(\lambda_2)$, and $T = X_1 + X_2$, recall we showed earlier that

$$\mathrm{E}(X_1 \mid T) = \frac{\lambda_1}{\lambda_1 + \lambda_2} T, \quad \mathrm{Var}(X_1 \mid T) = \frac{\lambda_1 \lambda_2}{(\lambda_1 + \lambda_2)^2} T$$

Recall $T \sim \text{Poisson}(\lambda_1 + \lambda_2)$ which implies $\mathrm{E}[T] = \lambda_1 + \lambda_2$ and $\mathrm{Var}(T) = \lambda_1 + \lambda_2$. It follows that

$$\mathrm{Var}(\mathrm{E}(X_1 \mid T)) = \mathrm{Var}\left(\frac{\lambda_1}{\lambda_1 + \lambda_2} T\right) = \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^2 \underbrace{\mathrm{Var}(T)}_{=\lambda_1 + \lambda_2} = \frac{\lambda_1^2}{\lambda_1 + \lambda_2}$$

$$\mathrm{E}[\mathrm{Var}(X_1 \mid T)] = \mathrm{E}\left[\frac{\lambda_1 \lambda_2}{(\lambda_1 + \lambda_2)^2} T\right] = \frac{\lambda_1 \lambda_2}{(\lambda_1 + \lambda_2)^2} \underbrace{\mathrm{E}[T]}_{=\lambda_1 + \lambda_2} = \frac{\lambda_1 \lambda_2}{\lambda_1 + \lambda_2}$$

Adding them up, we get

$$\mathrm{Var}(\mathrm{E}(X_1 \mid T)) + \mathrm{E}[\mathrm{Var}(X_1 \mid T)] = \frac{\lambda_1^2}{\lambda_1 + \lambda_2} + \frac{\lambda_1 \lambda_2}{\lambda_1 + \lambda_2} = \lambda_1 = \mathrm{Var}(X_1).$$

Tower Law of Variance is valid for this example.

Proof of Tower Law of Variance

By the shortcut formula for conditional variance,

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

taking expectation on both sides, we get

$$\mathrm{E}\left[\mathrm{Var}(Y \mid X)\right] = \underbrace{\mathrm{E}[\mathrm{E}(Y^2 \mid X)]}_{=\mathrm{E}[Y^2] \text{ by Tower Law}} - \mathrm{E}\left\{[\mathrm{E}(Y \mid X)]^2\right\}.$$

Proof of Tower Law of Variance

By the shortcut formula for conditional variance,

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

taking expectation on both sides, we get

$$\mathrm{E}\left[\mathrm{Var}(Y \mid X)\right] = \underbrace{\mathrm{E}[\mathrm{E}(Y^2 \mid X)]}_{=\mathrm{E}[Y^2] \text{ by Tower Law}} - \mathrm{E}\left\{[\mathrm{E}(Y \mid X)]^2\right\}.$$

Applying the shortcut formula for variance to $g(X) = \mathrm{E}(Y \mid X)$, we get

$$\mathrm{Var}(g(X)) = \mathrm{E}\left\{[g(X)]^2\right\} - (\mathrm{E}[g(X)])^2$$

$$\overset{g(X)=\mathrm{E}(Y|X)}{\Longrightarrow} \quad \mathrm{Var}(\mathrm{E}(Y \mid X)) = \mathrm{E}\left\{[\mathrm{E}(Y \mid X)]^2\right\} - \Big(\underbrace{\mathrm{E}[\mathrm{E}(Y \mid X)]}_{=\mathrm{E}[Y] \text{ by Tower Law}}\Big)^2$$

Proof of Tower Law of Variance

By the shortcut formula for conditional variance,

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

taking expectation on both sides, we get

$$\mathrm{E}\left[\mathrm{Var}(Y \mid X)\right] = \underbrace{\mathrm{E}[\mathrm{E}(Y^2 \mid X)]}_{=\mathrm{E}[Y^2] \text{ by Tower Law}} - \mathrm{E}\left\{[\mathrm{E}(Y \mid X)]^2\right\}.$$

Applying the shortcut formula for variance to $g(X) = \mathrm{E}(Y \mid X)$, we get

$$\mathrm{Var}(g(X)) = \mathrm{E}\left\{[g(X)]^2\right\} - (\mathrm{E}[g(X)])^2$$

$$\overset{g(X)=\mathrm{E}(Y|X)}{\Longrightarrow} \quad \mathrm{Var}(\mathrm{E}(Y \mid X)) = \mathrm{E}\left\{[\mathrm{E}(Y \mid X)]^2\right\} - \Bigg(\underbrace{\mathrm{E}[\mathrm{E}(Y \mid X)]}_{=\mathrm{E}[Y] \text{ by Tower Law}} \Bigg)^2$$

Adding them up, we get

$$\begin{aligned}\mathrm{E}\left[\mathrm{Var}(Y \mid X)\right] + \mathrm{Var}(\mathrm{E}(Y \mid X)) &= \mathrm{E}[Y^2] - \cancel{\mathrm{E}\left\{[\mathrm{E}(Y \mid X)]^2\right\}} + \cancel{\mathrm{E}\left\{[\mathrm{E}(Y \mid X)]^2\right\}} - (\mathrm{E}(Y))^2 \\ &= \mathrm{E}[Y^2] - (\mathrm{E}(Y))^2 = \mathrm{Var}(Y)\end{aligned}$$

Variance of Sum of a Random Number of R.V.'s

For a sum of the form

$$T = \sum_{i=1}^{N} X_i, \quad \text{where } \begin{cases} X_i's \text{ are i.i.d. with mean } \mathrm{E}(X) \\ N \text{ is indep of } X_i\text{'s} \end{cases}$$

We found earlier that $\mathrm{E}[T|N] = N\,\mathrm{E}(X)$.

$$\begin{aligned} \mathrm{Var}(T \mid N = n) &= \mathrm{Var}\left[\sum\nolimits_{i=1}^{n} X_i \,\middle|\, N = n\right] \\ &= \sum_{i=1}^{n} \mathrm{Var}(X_i \mid N = n) \quad \text{(as } X_i\text{'s are indep)} \\ &= \sum_{i=1}^{n} \underbrace{\mathrm{Var}\left[X_i \mid N = n\right]}_{\substack{= \mathrm{Var}(X) \text{ by indep.} \\ \text{of } N \text{ and } X_i\text{'s}}} = n\,\mathrm{Var}(X) \end{aligned}$$

This shows $\mathrm{Var}(T \mid N) = N\,\mathrm{Var}(X)$.

From $\mathrm{E}[T|N] = N\,\mathrm{E}(X)$ and $\mathrm{Var}(T \mid N) = N\,\mathrm{Var}(X)$, using the Tower Law for Variance,

$$\mathrm{E}[\mathrm{Var}(T \mid N)] = \mathrm{E}[N \overbrace{\mathrm{Var}(X)}^{\text{constant}}] = \mathrm{E}[N]\,\mathrm{Var}(X)$$
$$\mathrm{Var}(\mathrm{E}[T \mid N]) = \mathrm{Var}(N \underbrace{\mathrm{E}(X)}_{\text{constant}}) = (\mathrm{E}(X))^2\,\mathrm{Var}(N)$$

we get that

$$\mathrm{Var}(T) = \mathrm{Var}\left(\sum_{i=1}^{N} X_i\right) = \mathrm{E}[N]\,\mathrm{Var}(X) + (\mathrm{E}(X))^2\,\mathrm{Var}(N).$$

Example (Insurance Claims)

Suppose

- $N = \#$ of claims an insurance company receives in a month, $\sim$ Poisson(λ),
- the amounts of the individual claims $X_1, X_2, \ldots$ are i.i.d. with mean μ and variance σ^2.

The total amount of claims in the month is $T = \sum_{i=1}^{N} X_i$.

- $\mathrm{E}[T] = \mathrm{E}[N]\,\mathrm{E}[X] = \lambda\mu$
- $\mathrm{Var}[T] = \mathrm{E}[N]\,\mathrm{Var}(X) + (\mathrm{E}(X))^2\,\mathrm{Var}(N) = \lambda\sigma^2 + \mu^2\lambda$

Example (Mouse Trapped in a Maze) — Variance

Recall the total amount of time until escape is

$$T = \sum_{i=1}^{N} X_i, \text{ where } \begin{array}{l} X_i = \text{ travel time (in mins) of } i\text{th departure,} \\ N = \#\text{ of departures until free } \sim \text{Geometric}(p = 1/3) \end{array}$$

Clearly,

$$\mathrm{E}[N] = \frac{1}{p} = 3, \quad \mathrm{Var}(N) = \frac{1-p}{p^2} = 6.$$

To find $\mathrm{Var}(T \mid N = n)$:

- $X_N = 6$, always $\Rightarrow \mathrm{Var}(X_N) = 0$
- For $i < N$, X_i is equally likely to be 4 or 2 $\Rightarrow \mathrm{E}(X \mid N = n) = (4+2)/2 = 3$ and

$$\mathrm{Var}[X_i \mid N = n] = \frac{(4-3)^2 + (2-3)^2}{2} = 1, \quad i < n.$$

$$\text{So } \operatorname{Var}[T \mid N = n] = \operatorname{Var}\left[\sum_{i=1}^{n} X_i \,\middle|\, N = n\right]$$
$$= \sum_{i=1}^{n} \operatorname{Var}(X_i \mid N = n) \quad \text{(as } X_i\text{'s are indep)}$$
$$= \sum_{i=1}^{n-1} \underbrace{\operatorname{Var}(X_i \mid N = n)}_{=1} + \underbrace{\operatorname{Var}(X_n \mid N = n]}_{=0}$$
$$= n - 1$$

From $\operatorname{E}[T \mid N] = 3N + 3$ and $\operatorname{Var}[T \mid N] = N - 1$, using the Tower Law for Variance,

$$\operatorname{E}[\operatorname{Var}(T \mid N)] = \operatorname{E}[N - 1] = \operatorname{E}[N] - 1 = 3 - 1 = 2$$
$$\operatorname{Var}(\operatorname{E}[T \mid N]) = \operatorname{Var}(3N + 3) = 3^2 \operatorname{Var}(N) = 3^2 \cdot 6 = 54$$

we get that

$$\operatorname{Var}(T) = \operatorname{E}[\operatorname{Var}(T \mid N)] + \operatorname{Var}(\operatorname{E}[T \mid N]) = 2 + 54 = 56$$

On average, it takes the mouse 12 minutes to escape, give or take SD $= \sqrt{56} \approx 7.48$ minutes.

4.4.2 Prediction and Conditional Expectation

Predicting a Random variable Y by a Constant

How to best predict a random variable Y by a constant c?

- We want the predicted value c to be close to Y. A reasonable criterion would be to

$$\text{find c that minimize } \mathrm{E}\left[(Y-c)^2\right].$$

- The shortcut formula for $\mathrm{Var}(Y-c)$ gives

$$\underbrace{\mathrm{Var}(Y-c)}_{=\mathrm{Var}(Y)} = \mathrm{E}[(Y-c)^2] - \underbrace{(\mathrm{E}[Y-c])^2}_{=(\mathrm{E}(Y)-c)^2}.$$

Rearranging the terms, we get

$$\mathrm{E}[(Y-c)^2] = \mathrm{Var}(Y) + (\mathrm{E}(Y)-c)^2$$

This means that $\mathrm{E}[(Y-c)^2]$ is minimized when $c = \mathrm{E}[Y]$.

Prediction and Conditional Expectation

For two random variables X, Y with some joint distribution, if X is observed to be x, what's the best predicted value for Y?

Prediction and Conditional Expectation

For two random variables X, Y with some joint distribution, if X is observed to be x, what's the best predicted value for Y?

- The predicted value would depend on the observed X and hence must be a function $g(X)$ of X
- We want the predicted value $g(X)$ to be close to Y. A reasonable criterion would be to

$$\text{find } g(X) \text{ that minimize } \mathrm{E}\left[(Y - g(X))^2 \,\middle|\, X\right].$$

- As $\mathrm{E}\left[(Y - c)^2\right]$ is minimized when $c = \mathrm{E}(Y)$, similarly, $\mathrm{E}\left[(Y - g(X))^2 \mid X\right]$ is minimized when

$$g(X) = \mathrm{E}[Y \mid X].$$

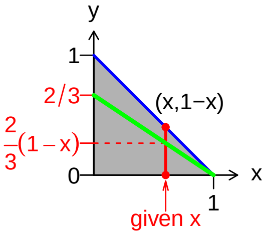
Example (Mixed Nuts) — Prediction

Recall in L05, the joint PDF for

$$X = \text{ the weight of almonds, and}$$
$$Y = \text{ the weight of cashews}$$

in a can of mixed nuts is

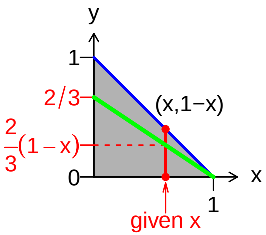
$$f(x,y) = \begin{cases} 24xy & \text{if } 0 \le x, y \le 1, x + y < 1 \\ 0 & \text{otherwise} \end{cases}$$



Example (Mixed Nuts) — Prediction

Recall in L05, the joint PDF for

$$X = \text{ the weight of almonds, and}$$
$$Y = \text{ the weight of cashews}$$



in a can of mixed nuts is

$$f(x,y) = \begin{cases} 24xy & \text{if } 0 \le x, y \le 1, x + y < 1 \\ 0 & \text{otherwise} \end{cases}$$

We showed earlier that $\mathrm{E}(Y \mid X) = \frac{2}{3}(1 - X)$.

Given there were $X = x$ lbs of almonds in a can, our best prediction for the amount of cashews in the can is

$$\mathrm{E}(Y \mid X) = \frac{2}{3}(1 - X) \text{ lbs.}$$