Here I hope to summarize in convenient form the facts and formulas associated with conditional expectation of random variables given various things — events, other random variables' values, just plain other random variables, etc. One recurring theme is the idea that pmfs and pdfs have means and variances. We think of means and variances as things random variables have, but any legitimate pmf or pdf has a mean and a variance, where by "legitimate pmf" I mean that it's nonnegative and sums over x to 1, and by "legitimate pdf" I mean that it's nonnegative and integrates over x to 1. Specifically.

- Any legitimate pmf $p(x)$ has mean and variance
$$\text{mean} = \sum_x xp(x) \qquad \text{variance} = \sum_x (x - \text{mean})^2 p(x) \ .$$
- Any legitimate conditional pmf $p(x \mid \text{stuff})$ has mean and variance
$$\text{mean} = \sum_x xp(x \mid \text{stuff}) \qquad \text{variance} = \sum_x (x - \text{mean})^2 p(x \mid \text{stuff}) \ .$$
- Any legitimate pdf $f(x)$ has mean and variance a mean
$$\text{mean} = \int_{-\infty}^{\infty} xf(x)dx \qquad \text{variance} = \int_{-\infty}^{\infty} (x - \text{mean})^2 f(x)dx \ .$$
- Any legitimate conditional pdf $f(x \mid \text{stuff})$ has mean and variance
$$\text{mean} = \int_{-\infty}^{\infty} xf(x \mid \text{stuff})dx \qquad \text{variance} = \int_{-\infty}^{\infty} (x - \text{mean})^2 f(x \mid \text{stuff})dx \ .$$

In what follows, I'll refer often to means and variances of various conditional pmfs and pdfs.

First let's talk about conditioning on events. If A is an event with positive probability and X is a discrete random variable, the conditional pmf of X given A is

$$p_{X|A}(x) = \frac{\mathbb{P}(\{X = x\} \cap A)}{\mathbb{P}(A)} \ .$$

This is a legitimate pmf as a function of x. Its mean is

$$\mathbb{E}(X \mid A) = \sum_x xp_{X|A}(x)$$

and its variance is

$$\text{Var}(X \mid A) = \sum_x (x - \mathbb{E}(X \mid A))^2 p_{X|A}(x) \ .$$

If X is instead a continuous random variable , the conditional pdf of X given A exists, but we don't generally have a nice formula for it. You can always calculate the conditional cdf of X given A, which is

$$F_{X|A}(x) = \frac{\mathbb{P}(\{X \le x\} \cap A)}{\mathbb{P}(A)}$$

and take the derivative to get

$$f_{X|A}(x) = \frac{d}{dx} F_{X|A}(x) \ .$$

In the special case where A is an event of the form $\{X \in W\}$, it turns out that

$$f_{X|A}(x) = \begin{cases} \frac{f_X(x)}{\mathbb{P}(A)} & \text{when } x \in W \\ 0 & \text{when } x \notin W \ . \end{cases}$$

In any case, $f_{X|A}$ is a legitimate pdf as a function of x and has a mean

$$\mathbb{E}(X \mid A) = \int_{-\infty}^{\infty} xf_{X|A}(x)dx$$

and variance

$$\mathrm{Var}(X \mid A) = \int_{-\infty}^{\infty} (x - \mathbb{E}(X \mid A))^2 f_{X|A}(x)dx \; .$$

If A_k, $1 \leq k \leq n$, are events that partition the sample space, then in either case — X discrete or X continuous — the law of total expectation states that

$$\mathbb{E}(X) = \sum_{k=1}^{n} \mathbb{E}(X \mid A_k)\mathbb{P}(A_k) \; .$$

Now let's talk about conditioning on other random variables. Given random variables X and Y, four distinct situations arise. Both random variables can be discrete, both continuous, or one discrete and the other continuous.

- **Both discrete:** The conditional pmf of X given $Y = y$ is

  $$p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x, y)}{p_Y(y)} \; .$$

  We derived this formula by starting with conditional pmfs of X given events of the form $\{Y = y\}$. This is a legitimate pmf as a function of x and has mean

  $$\mathbb{E}(X \mid Y = y) = \sum_x x p_{X|Y}(x \mid y)$$

  and variance

  $$\mathrm{Var}(X \mid Y = y) = \sum_x (x - \mathbb{E}(X \mid Y = y))^2 p_{X|Y}(x \mid y) \; .$$

  The law of total expectation states that

  $$\mathbb{E}(X) = \sum_y \mathbb{E}(X \mid Y = y) p_Y(y) \; .$$

- **Both continuous:** The conditional pdf of X given $Y = y$ is

  $$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \; .$$

  We didn't derive this formula directly by starting with conditional pdfs of X given events of the form $\{Y = y\}$ because those events have zero probability but instead employed a limiting process. $f_{X|Y}(x \mid y)$ is a legitimate pdf as a function of x and has mean

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

  and variance

  $$\mathrm{Var}(X \mid Y = y) = \int_{-\infty}^{\infty} (x - \mathbb{E}(X \mid Y = y))^2 f_{X|Y}(x \mid y)dx \; .$$

  The law of total expectation states that

  $$\mathbb{E}(X) = \int_{-\infty}^{\infty} \mathbb{E}(X \mid Y = y) f_Y(y)dy \; .$$

- X **continuous,** Y **discrete:** The conditional pdf of X given $Y = y$ is

  $$f_{X|Y}(x \mid y) = f_{X|A_y}(x) \; ,$$

  where A_y is the event $\{Y = y\}$. This is a legitimate pdf as a function of x and has mean

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

  and variance

  $$\mathrm{Var}(X \mid Y = y) = \int_{-\infty}^{\infty} (x - \mathbb{E}(X \mid Y = y))^2 f_{X|Y}(x \mid y)dx \; .$$

The law of total expectation states that

$$\mathbb{E}(X) = \sum_y \mathbb{E}(X \mid Y = y) p_Y(y) \; .$$

- X **discrete,** Y **continuous:** Here, we don't generally have a nice formula for the conditional pmf of X given $Y = y$. We can define it officially through a limiting process

$$p_{X|Y}(x \mid y) = \lim_{\delta \to 0} \frac{\mathbb{P}(\{X = x\} \cap \{Y \in [y - \delta, y + \delta]\})}{\mathbb{P}(\{Y \in [y - \delta, y + \delta])} \; ,$$

but typically we don't have to go that route. In many settings, we have at hand the conditional pdfs $f_{Y|X}(y \mid x)$ for all values of x along with the marginal pmf $p_X(x)$. From these we can, by a limiting process, derive the formula

$$p_{X|Y}(x \mid y) = \frac{f_{Y|X}(y \mid x) p_X(x)}{f_Y(y)} = \frac{f_{Y|X}(y \mid x) p_X(x)}{\sum_x f_{Y|X}(y \mid x) p_X(x)} \; ,$$

where the second equality holds because of the law of total probability. In any case, $p_{X|Y}(x \mid y)$ is a legitimate pmf as a function of X and has mean

$$\mathbb{E}(X \mid Y = y) = \sum_x x p_{X|Y}(x \mid y)$$

and variance

$$\mathrm{Var}(X \mid Y = y) = \sum_x (x - \mathbb{E}(X \mid Y = y))^2 p_{X|Y}(x \mid y) \; .$$

The law of total expectation states that

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} \mathbb{E}(X \mid Y = y) f_Y(y) dy \; .$$

In all four situations outlined above $\mathbb{E}(X \mid Y = y)$ is a number for each y. Let's call that number $g(y)$. If we plug Y into the function g, we get $g(Y)$, which is a random variable. That random variable we call $\mathbb{E}(X \mid Y)$. Note that $\mathbb{E}(X \mid Y)$ is discrete when Y is discrete and may be continuous only when when Y is continuous, but $g(Y)$ might be discrete even when Y is continuous. As a random variable, it has an expected value. By the expected value rule,

$$\begin{aligned}
\mathbb{E}(g(Y)) &= \begin{cases} \sum_y g(y) p_Y(y) & \text{when } Y \text{ is discrete} \\ \int_{-\infty}^{\infty} g(y) f_Y(y) & \text{when } Y \text{ is continuous} \end{cases} \\
&= \begin{cases} \sum_y \mathbb{E}(X \mid Y = y) p_Y(y) & \text{when } Y \text{ is discrete} \\ \int_{-\infty}^{\infty} \mathbb{E}(X \mid Y = y) f_Y(y) & \text{when } Y \text{ is continuous} \end{cases} \\
&= \mathbb{E}(X) \; ,
\end{aligned}$$

where the last equality follows from the law of total expectation featured in the four bulleted items above. Thus we have the law of iterated expectations, namely

$$\mathbb{E}(\mathbb{E}(X \mid Y)) = \mathbb{E}(X) \; ,$$

which holds no matter what kind of random variables X and Y are.

Furthermore, in all four situations outlined above $\mathrm{Var}(X \mid Y = y)$ is a number for each y. Let's call that number $\gamma(y)$. If we plug Y into the function γ, we get $\gamma(Y)$, which is a random variable. That random variable we call $\mathrm{Var}(X \mid Y)$. Note that $\mathrm{Var}(X \mid Y)$, like $\mathbb{E}(X \mid Y)$, is discrete when Y is discrete and may be continuous only when Y is continuous. We can also define $\mathrm{Var}(X \mid Y)$ by the formula

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

To see why this works, let $g(y) = \mathbb{E}(X \mid Y = y)$, so $\mathbb{E}(X \mid Y) = g(Y)$. Then, assuming for convenience that X is discrete, we get

$$\begin{aligned}
\psi(y) &= \mathbb{E}\left((X - \mathbb{E}(X \mid Y))^2 \mid Y = y\right) \\
&= \mathbb{E}\left((X - g(Y))^2 \mid Y = y\right) \\
&= \mathbb{E}\left((X - g(y))^2 \mid Y = y\right) \\
&= \sum_x (x - g(y))^2 p_{X|Y}(x \mid y) \\
&= \mathrm{Var}(X \mid Y = y) \\
&= \gamma(y) \; .
\end{aligned}$$

Plugging Y in for y in $\psi(y)$ gives us the random variable $\mathbb{E}\left((X - \mathbb{E}(X \mid Y))^2 \mid Y\right)$, which is the same as $\gamma(Y)$ because $\psi(y) = \gamma(y)$ for all y. A similar argument works when X is continuous.

As for ordinary variance, we have a formula for $\mathrm{Var}(X \mid Y)$ in terms of moments, namely

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

This follows from the second definition for conditional variance because

$$\begin{aligned}
\mathrm{Var}(X \mid Y) &= \mathbb{E}\left((X - \mathbb{E}(X \mid Y))^2 \mid Y\right) \\
&= \mathbb{E}\left(X^2 \mid Y\right) - 2\mathbb{E}(X\mathbb{E}(X \mid Y)) + \mathbb{E}\left((\mathbb{E}(X \mid Y))^2 \mid Y\right) \\
&= \mathbb{E}\left(X^2 \mid Y\right) - 2\mathbb{E}(X \mid Y)\mathbb{E}(X \mid Y) + (\mathbb{E}(X \mid Y))^2 \\
&= \mathbb{E}\left(X^2\right) - (\mathbb{E}(X \mid Y))^2 \; ,
\end{aligned}$$

where the second-to-last line follows from the fact that $\mathbb{E}(X \mid Y)$ is a function of Y, so

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

and

$$\mathbb{E}\left((\mathbb{E}(X \mid Y))^2 \mid Y\right) = (\mathbb{E}(X \mid Y))^2 \; .$$

The law of total variance states that

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

This follows from the fact that $\mathbb{E}(X \mid Y)$ and $X - \mathbb{E}(X \mid Y)$ are uncorrelated and sum to X, so their variances sum to $\mathrm{Var}(X)$. Thus

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{Var}(X - \mathbb{E}(X \mid Y)) + \mathrm{Var}(\mathbb{E}(X \mid Y)) \\
&= \mathbb{E}\left((X - \mathbb{E}(X \mid Y))^2\right) + \mathrm{Var}(\mathbb{E}(X \mid Y)) \\
&= \mathbb{E}\left(\mathbb{E}\left((X - \mathbb{E}(X \mid Y))^2 \mid Y\right)\right) + \mathrm{Var}(\mathbb{E}(X \mid Y)) \\
&= \mathbb{E}(\mathrm{Var}(X \mid Y)) + \mathrm{Var}(\mathbb{E}(X \mid Y)) \; ,
\end{aligned}$$

where the second line holds because $X - \mathbb{E}(X \mid Y)$ has zero mean and the third because of the law of iterated expectations. and the last because $\mathrm{Var}(X \mid Y)$

Like conditional probability, conditional expectation arises in situations involving more than two random variables. Given three random variables X, Y, and Z defined on the same probability space, $\mathbb{E}(X \mid Y = y, Z = z)$ is a number for each y and z. Let's call that number $h(y, z)$, thus defining a function of y and z. Plug Y and Z in $h(y, z)$ for y and z and you have the random variable $\mathbb{E}(X \mid Y, Z)$, the conditional expectation of X given Y and Z. Things can get complicated when the three random variables constitute an assortment of discrete and continuous random variables, but let's see how this works when all of them are either discrete or continuous.

If all are discrete, for any y and z we have the conditional pmf

$$p_{X|Y,Z}(x \mid y, z) = \frac{p_{X,Y,Z}(x, y, z)}{p_{Y,Z}(y, z)} \; ,$$

which is a legitimate pmf as a function of X and has mean

$$h(y,z) = \mathbb{E}(X \mid Y = y, Z = z) = \sum_x x p_{X|Y,Z}(x \mid y, z) \ .$$

If all are continuous, then for any y and z we have the conditional pdf

$$f_{X|Y,Z}(x \mid y, z) = \frac{f_{X,Y,Z}(x,y,z)}{f_{Y,Z}(y,z)} \ ,$$

which is a legitimate pdf as a function of X and has mean

$$h(y,z) = \mathbb{E}(X \mid Y = y, Z = z) = \int_{-\infty}^{\infty} x f_{X|Y,Z}(x \mid y, z) dx \ .$$

In either case, $\mathbb{E}(X \mid Y, Z) = h(Y, Z)$.

The law of iterated expectations extends as follows:

$$\mathbb{E}(\mathbb{E}(X \mid Y, Z) \mid Z) = \mathbb{E}(X \mid Z) \ .$$

I'll prove it in the case when all the random variables are discrete, but it holds in general. First let $g(z) = \mathbb{E}(X \mid Z = z)$ and $h(y,z) = \mathbb{E}(X \mid Y, Z)$. Then for any z we have

$$\mathbb{E}(\mathbb{E}(X \mid Y, Z) \mid Z = z) = \mathbb{E}(h(Y,Z) \mid Z = z) = \mathbb{E}(h(Y,z) \mid Z = z) \ .$$

Note that $h(Y,z)$ appearing in the rightmost term is a function of Y, so by the expected value rule

$$\begin{aligned}
\mathbb{E}(h(Y,z) \mid Z = z) &= \sum_y h(y,z) p_{Y|Z}(y \mid z) \\
&= \sum_y \left(\sum_x x p_{X|Y,Z}(x \mid y, z) \right) p_{Y|Z}(y \mid z) \\
&= \sum_x x \left(\sum_y p_{X|Y,Z}(x \mid y, z) p_{Y|Z}(y \mid z) \right) \\
&= \sum_x x p_{X|Z}(x \mid z) \\
&= g(z) \ ,
\end{aligned}$$

where I used the identity

$$p_{X|Y,Z}(x \mid y, z) p_{Y|Z}(y \mid z) = \frac{p_{X,Y,Z}(x,y,z)}{p_{Y,Z}(y,z)} \frac{p_{Y,Z}(y,z)}{p_Z(z)} = \frac{p_{X,Y,Z}(x,y,z)}{p_Z(z)} \ ,$$

which sums over y to give $p_{X|Z}(x \mid z)$. The bottom line is that

$$\mathbb{E}(\mathbb{E}(X \mid Y, Z) \mid Z = z) = \mathbb{E}(X \mid Z = z) = g(z)$$

for all z, and the extended law of iterated expectations follows.

The foregoing extends naturally to cover things like

$$\mathbb{E}\left(X \mid Y_1, Y_2, \ldots, Y_n\right) \ ,$$

and various laws of iterated expectations hold, all expressible as

$$\mathbb{E}(\mathbb{E}(X \mid \text{list}) \mid \text{sublist}) = \mathbb{E}(X \mid \text{sublist}) \ ,$$

where "list" represents a list of random variables and "sublist" represents any sublist of "list." These extended laws of iterated expectations make particular sense when you think about conditional expectation as a projection operation, which I like to do. Let me explain.

In class we showed that $\mathbb{E}(X \mid Y)$ minimizes $\mathbb{E}\left((X - h(Y))^2\right)$ over all functions $h(Y)$. The picture Alex generated to accompany Rami's notes illustrates $\mathbb{E}(X \mid Y)$ as the orthogonal projection of X onto the space of all random variables of the form $h(Y)$, where orthogonal means uncorrelated, as the error $X - \mathbb{E}(X \mid Y)$ is with all functions $h(Y)$.

Similarly, you can show that $\mathbb{E}(X \mid Y, Z)$ minimizes $\mathbb{E}\left((X - h(Y, Z))^2\right)$ over all functions $h(Y, Z)$, and it amounts essentially to the orthogonal projection of X on to the space of all such functions.

Now, when you project X orthogonally onto a big space and in turn project the projection you get onto a subspace of the big space, you get get the projection of X onto the subspace. In the present context, $\mathbb{E}(X \mid Y, Z)$ is the projection X onto the space of functions of Y and Z, and $\mathbb{E}(\mathbb{E}(X \mid Y, Z) \mid Z)$ is the projection of that projection onto the space of all functions of Z, which is s aubspace of the space of all functions of Y and Z. That last projection is $\mathbb{E}(X \mid Z)$, which is what you would have gotten had you just projected X onto the space of functions of Z to start with.