Because we didn't get to talk at too much length about the formal definition of conditional expected values, I thought it worth jotting down a few points for you to read. These notes are quick and rough, but please go through them and come talk to me if you have questions.

To start with, assume we have a base probability space $(\Omega, \mathcal{F}, \mathsf{E})$ that is rich enough to support any model we care to construct. Just as a reminder, remember what a probability space is. Ω is the outcome space, each element of which – an elementary outcome $\omega \in \Omega$ – describes one way the random experiment comes out. The σ-field $\mathcal{F}$ is a collection of subsets of Ω, called events, containing those sets whose indicators are measurable functions. And E is the expected value operator, which must satisfy certain axioms but which is otherwise free to be set by our model. We could equivalently define the probability operator P, where $\mathsf{P}(A) = \mathsf{E}1_A$. (In fact, some texts use P to denote both the expected value operator on random variables and the set function that gives expected values of the corresponding integers.)

Now the first point I want to make is the intuition I described in class behind using a σ-field of events to represent information we have at some point during an experiment.

Intuition 1. Let $\mathcal{G} \subset \mathcal{F}$ be a σ-field. To say that we have information $\mathcal{G}$ means that we know the value of every indicator 1_G for $G \in \mathcal{G}$. Or to put it another way, the information $\mathcal{G}$ tells us whether each event $G \in \mathcal{G}$ occurred.

Example 2. Let $\mathcal{G} = \{\emptyset, \Omega\}$. What information is embodied in that σ-field?

"Knowing" $\mathcal{G}$ corresponds to knowing the value of 1_G for every $G \in \mathcal{G}$ (that is, knowing whether each event G occurred, or not). But $1_\emptyset \equiv 0$ and $1_\Omega \equiv 1$, so we knew this information already.

In other words, having the information $\mathcal{G}$ means having no extra information at all.

Example 3. X denote the value of a six-sided die. Let $\mathcal{G}$ be the smallest σ-field containing the events $G_1 = \{X = 1\}$, $G_2 = \{X = 2\}$, ..., $G_6 = \{X = 6\}$. (What is that σ-field?)

"Knowing" the information in $\mathcal{G}$ means, for example, that we know the values of 1_{G_k} for $k = 1, \ldots, 6$. And in particular, we know the value of

$$\sum_{k=1}^{6} k 1_{G_k}.$$

But this random variable is just X! Knowing the information in $\mathcal{G}$ tells us the value of X. And conversely, if I know the value of X, I know 1_G for all $G \in \mathcal{G}$ because *all these events are events that relate to the value of* X.

In this sense, the σ-field $\mathcal{G}$ embodies the information in X!

Example 4. Taking this to greater generality, let X now be a real-valued random variable with $X \geq 0$.

Let $\mathcal{G}$ be the smallest σ-field generated by the events $\{X \in A\}$ for all A in the (extended) Borel σ-field on $\mathbb{R}$.

It turns out (see the Measures handout from earlier in the semester for the argument) that we can find random variables $X_1 \leq X_2 \leq X_3 \leq \cdots$ such that

1. $X_n \nearrow X$ (that is, $\lim_n X_n = X$ and the X_ns are increasing)., and
2. Each X_n is of the form

$$X_n = \sum_{k=1}^{n} c_{nk} 1_{G_{nk}}$$

for constants c_{nk} and events $G_{nk} \in \mathcal{G}$. (Such a function is called a *simple* function because it takes only a finite number of non-zero values.) (It's not necessary to make the number of terms in X_n exactly n, any finite number will do, but we can so what the heck.)

Now, "knowing" the information in $\mathcal{G}$ means that we know the value of every X_n. And this in turn means that we know the value of X.

Conversely, if we know the value of X, we know 1_G for each $G \in \mathcal{G}$ because each G is generated from events that are based on the value of X.

Hence again, the σ-field $\mathcal{G}$ embodies all the infromation in X.

This example generalizes to real-valued random variables and beyond. I won't do the argument here, but see the measures handout; it's the same construction used in defining the integral.

Definition 5. For a random variable X (vector or scalar valued), we define $\sigma(X)$ to be the minimal σ-field generated by events of the form $\{X \in A\}$ for A in the (extended) σ-field in the range measurable space. Given a collection of random variables $X_1, X_2, \ldots, X_n$, we write $\sigma(X_1, \ldots, X_n)$ interchangeably with $\sigma(X)$ for $X = (X_1, \ldots, X_n)$. This works for infinite collections too, but we have to be more careful in defining the σ-field. More on that another time.

Reminder 6. Recall that if $f\colon (\mathcal{X}, \mathcal{A}) \to (\mathcal{Y}, \mathcal{B})$ is a mapping between measurable spaces, then f is a measurable function if $f^{-1}(B) \in \mathcal{A}$ for each $B \in \mathcal{B}$. If $\mathcal{C} \subset \mathcal{A}$ is a σ-field, then f is $\mathcal{C}$-measurable if $f^{-1}(B) \in \mathcal{C}$ for each $B \in \mathcal{B}$.

Now we are in a position to consider the formal definition again.

Formal Definition (Part I) 7. Let $(\Omega, \mathcal{F}, \mathsf{E})$ be a probability space and let $\mathcal{G} \subset \mathcal{F}$ be a σ-field. If Y is a random variable such that $\mathsf{E}Y$ exists, then there exists a random variable, denoted by $\mathsf{E}(Y \mid \mathcal{G})$, with the following properties:

1. $\mathsf{E}(Y \mid \mathcal{G})$ is $\mathcal{G}$-measurable.
2. $\mathsf{E}\left(\mathsf{E}(Y \mid \mathcal{G})\right)$ exists.
3. For every $G \in \mathcal{G}$,
$$\mathsf{E}\left(\mathsf{E}(Y \mid \mathcal{G})1_G\right) = \mathsf{E}\left(Y 1_G\right), \tag{1}$$
or equivalently
$$\mathsf{E}1_G\left(Y - \mathsf{E}(Y \mid \mathcal{G})\right) = 0. \tag{2}$$

This extends naturally to vector-valued variables

(Rigor alert: the random variable $\mathsf{E}(Y \mid \mathcal{G})$ is *not* unique, but it is unique up to events of probability zero. Moreover, we can find one version of this random variable with all the regularity properties we would expect, just i's and t's, folks.)

The intuition behind this definition is quite lovely. Keep in mind the notion from above of a σ-field representing information we might learn during an experiment.

What does it mean for a $\mathcal{G}$-measurable random variable Z to be $\mathsf{E}(Y \mid \mathcal{G})$? (I'm using Z here because it's easier to think about that as a random variable, but it's just a version – unique up to a set of probability zero – of the conditional expected value.) If $G \in \mathcal{G}$ has occurred (and assume $\mathsf{E}1_G > 0$), then we'd like to know that Y and Z will have the same long run average over G. That is, we want

$$\frac{\mathsf{E}Y1_G}{\mathsf{E}1_G} = \frac{\mathsf{E}Z1_G}{\mathsf{E}1_G} \iff \mathsf{E}Y1_G = \mathsf{E}Z1_G. \tag{3}$$

Since $\mathcal{G}$ is a σ-field, G^c is also in $\mathcal{G}$, so the same argument requires that

$$\frac{\mathsf{E}Y(1-1_G)}{\mathsf{E}(1-1_G)} = \frac{\mathsf{E}Z(1-1_G)}{\mathsf{E}(1-1_G)} \iff \mathsf{E}Y(1-1_G) = \mathsf{E}Z(1-1_G). \tag{4}$$

The left equations in the above two equivalences should look familiar; they are the values of the predictors of Y or Z given 1_G. That is, we require:

$$\mathsf{pred}_{Y|1_G} = \mathsf{pred}_{Z|1_G}. \tag{5}$$

This is an equality of functions, meaning that the *predictors are the same*. That is, given 1_G, your mean-square optimal predictor of Y is also your mean-square optimal predictor of Z.

Looking at the right size of the above equivalences, we can rewrite the first as

$$\mathsf{E}1_G(Y - Z) = 0. \tag{6}$$

This says that the difference between Y and Z – the residual we might call it – averages to zero over $G \in \mathcal{G}$. The analogy to residuals in list of numbers that I gave in class is a useful one. If Y were, say, greater than Z over some set in $\mathcal{G}$, this condition would not hold. To put it loosely, the residual is, relative to $\mathcal{G}$, noise. And in fact this is what suggests that Z

is an optimal $\mathcal{G}$-measurable predictor of Y, using basically the arguments we've seen in class and earlier in this document.

To see what this means further, note that assuming that $\mathsf{E}Y$ and $\mathsf{E}Z$ exist, if $0 \leq X_1 \leq X_2 \leq \cdots$ are $\mathcal{G}$-measurable random variables that converge to some $\mathcal{G}$-measurable X, then $\mathsf{E}X_n|Y-Z|$ increases to $\mathsf{E}X|Y-Z|$ by the Monotone Limits Rule and so by a related theorem $\mathsf{E}X_n(Y-Z) \to \mathsf{E}X(Y-Z)$. If X is non-negative $\mathcal{G}$ measureable functions and the X_n are simple $\mathcal{G}$-measurable functions increasing to X as used earlier, then linearity of expected values and equation (6) show that $\mathsf{E}X_n(Y-Z) = 0$ for all n. Thus, $\mathsf{E}X(Y-Z)$. For an arbitrary real-valued $\mathcal{G}$ measurable function, we can write $X = \max(X,0) - \min(X_0) \equiv X_+ - X_-$ and apply the same argument to X_+ and X_-. The bottom line is that for any $\mathcal{G}$-measurable function $\mathsf{E}X(Y-Z) = 0$.

Now, consider an arbitrary $\mathcal{G}$-measurable random variable U

$$\mathsf{E}(Y-U)^2 = \mathsf{E}(Y-Z+Z-U)^2 = \mathsf{E}(Y-Z)^2 + 2\mathsf{E}(Z-U)(Y-Z) + \mathsf{E}(Z-U)^2. \quad (7)$$

Look at that! The first term does not depend on U. The second term is zero by our condition equation (6) (make sure you understand why that is so!). The third term is quite immediately minimized when $U = Z$. Hence, our Z is just the optimal mean-squared $\mathcal{G}$-measurable predictor of Y.

We thus have three equivalent requirements for a random variable Z that purports to be $\mathsf{E}(Y \mid \mathcal{G})$.

1. The long-run average of Y and Z over events in $\mathcal{G}$ must be the same. ($\mathsf{E}Y1_G = \mathsf{E}Z1_G$)
2. The best predictors of Y and Z given information in $\mathcal{G}$ are the same.
3. The residual $(Y-Z)$ averages to zero over $G \in \mathcal{G}$. ($\mathsf{E}1_G(Y-Z) = 0$). This implies, in turn that Z is the optimal mean-square predictor.

Now, let's unpack the formal definition. First, the random variable $\mathsf{E}(Y \mid \mathcal{G})$ exists. Good thing. Second, it's $\mathcal{G}$-measurable. That's the way we cast it when we talked about Z above being $\mathcal{G}$-measurable. Third, $\mathsf{E}\mathsf{E}(Y \mid \mathcal{G})$ exists. A technical requirement but necessary if this is going to be useful. Fourth, $\mathsf{E}Y1_G = \mathsf{E}\mathsf{E}(Y \mid \mathcal{G})1_G$; that's just what we discussed, pick your favorite interpretation from the above 3. The third in particular connects to the heuristic definitions which will prove so useful.

You may be wondering next about the following.

Formal Definition (Part II) 8. Assume the conditions of the previous definition. Let X be a random variable (real or vector-valued) and define $\mathcal{G} = \{X^{-1}(A)\}$ for sets in the corresponding σ-field in the range of X. Then $\mathcal{G}$ is a σ-field and we define

$$\mathsf{E}(Y \mid X) = \mathsf{E}(Y \mid \mathcal{G}), \quad (8)$$

where we can assume we've picked a "nice" version. The defining property of the conditional expectation corresponds to the following useful identity for any (measurable) function g:

$$\mathsf{E}\left(\mathsf{E}(Y \mid X)g(X)\right) = \mathsf{E}\left(Yg(X)\right), \quad (9)$$

or more compellingly

$$\mathsf{E}g(X)(Y - \mathsf{E}(Y \mid X)) = 0. \quad (10)$$

Think back to the examples above. The σ-field $\sigma(X)$ generated by events $\{X \in \mathcal{A}\}$ embodies the information in X as described above. Hence, defining $\mathsf{E}(Y \mid X) = \mathsf{E}(Y \mid \sigma(X)) = \mathsf{E}(Y \mid \mathcal{G})$ makes sense; and note that it is also consistent with our prediction scheme. Hurray!