Measure Theoretic Conditional Expectation in an Elementary Setting

Author: *R. Scott McIntire*

Date: Mar 25, 2025

Overview

The measure theoretic approach to conditional expectation can be confusing when compared to the traditional approach. In what follows we go through a conditional expectation problem within a discrete and familiar setting in an attempt to reduce this confusion. One of the most confusing/disturbing ideas that one must confront from the very definition of conditional expectation is the notion of non-measurable functions. We work through an explicit example where we demonstrate that a given function is not measurable in the "conditional" σ-algebra.

Elementary Probability Example

Let $X = \{D_1, D_2, D_3, D_4, D_5, D_6\}$ and define a function P by $P(D_i) = \frac{1}{6}$, for $i \in \{1, 2, 3, 4, 5, 6\}$. The intent is that P will become a probability measure for the space we construct. Let $\mathcal{E} = 2^X$ be the σ-algebra consisting of the power set of X. We extend P for every element in the σ-algebra. Since the σ-algebra consists of all sets we need an assignment for an arbitrary set, A. The assignment is $P(A) = \frac{|A|}{6}$; that is, the cardinality of the set divided by 6. We now have a measure space; in fact, a probability space: $(P, X, \mathcal{E})$. Note that for a probability space we need an event space, X, a σ-algebra of sets (in the discrete case just an algebra), and a function P which takes elements of the σ-algebra to [0,1] with the following properties:

- $P(X) = 1$
- $P(\emptyset) = 0.$
- $P\left(\bigcup_{i=1}^{N} A_i\right) = \sum_{i=1}^{N} P(A_i) \quad \text{when } A_k \cap A_j = \emptyset \quad k \neq j;$

We now consider a random variable from which we will get a sub σ-algebra. Let

$$g(D_i) = \begin{cases} 0 & \text{if } i \text{ is even;} \\ 1 & \text{if } i \text{ is odd} \end{cases}$$

In a discrete space the σ-algebra generated from g is the algebra of sets generated from the sets: $g^{-1}(a)$, $a \in (-\infty, \infty)$. Clearly, the interesting sets come from the values 0 and 1, all other values lead to the empty set. Consequently,

$$\mathcal{F} = \{\emptyset, \{D_1, D_3, D_5\}, \{D_2, D_4, D_6\}, \{D_1, D_2, D_3, D_4, D_5, D_6\}\}$$

In a discrete space a function, f, is measurable with respect to a σ-algebra if $f^{-1}(a)$ is an element in the σ-algebra for all $a \in (-\infty, \infty)$. This has implications for the σ-algebra $\mathcal{F}$.

claim: Any function, f, which is measurable over $\mathcal{F}$ has the property that f is constant on the sets $\{D_1, D_3, D_5\}$ and $\{D_2, D_4, D_6\}$. More generally, we claim that any measurable function, f, in a given σ-algebra must be constant on the *minimal* elements of the algebra – elements which have no non-trivial subsets.†

To see this suppose that $f(1)$ differs from $f(3)$. Then the set, $A = \{D_1, D_3, D_5\} \cap f^{-1}(f(D_1))$, is a *strict*, non-trivial subset of $\{D_1, D_3, D_5\}$. This is true since A, by construction, is a subset of $\{D_1, D_3, D_5\}$; A contains D_1; and A does not contain D_3. Since $\mathcal{F}$ is a σ-algebra and A is the intersection of the $\mathcal{F}$ measurable sets $\{D_1, D_3, D_5\}$ and $f^{-1}(f(D_1))$, A must be in $\mathcal{F}$. However, we know that $\{D_1, D_2, D_3\}$ has no strict non-trivial subset in $\mathcal{F}$ – contradiction. Therefore, our premise that $f(D_1)$ and $f(D_3)$ could take differing values is incorrect. The same argument shows that $f(D_1)$ and $f(D_5)$ do not differ. One can repeat the above argument to show that f is constant on the other minimal set $\{D_2, D_4, D_6\}$.

Notice that while any function over the measure space $(P, X, \mathcal{E})$ is measurable, we can write down a specific function that is non-measurable with respect to $\mathcal{F}$. We know that all we have to do is come up with a function that differs on either of the sets $\{D_1, D_3, D_5\}$ or $\{D_2, D_4, D_6\}$. For instance, the function: $f(D_i) = i$, for $i \in \{1, 2, 3, 4, 5, 6\}$, is a non-measurable function in $\mathcal{F}$.

Conditional Expectation

Given a probability space $(P, X, \mathcal{E})$, the conditional expectation of a measurable function f with respect to a sub σ-algebra $\mathcal{F}$ is the *unique $\mathcal{F}$ measurable* function labeled, $\mathbb{E}[f/\mathcal{F}]$, such that‡

$$\int_\Lambda \mathbb{E}[f/\mathcal{F}]\; dP = \int_\Lambda f\, dP \quad \forall \Lambda \in \mathcal{F} \tag{1a}$$

Let us write this again in, perhaps, an unusual way.

$$\int_\Lambda \mathbb{E}[f/\mathcal{F}]\; dP_{\mathcal{F}} = \int_\Lambda f\, dP \quad \forall \Lambda \in \mathcal{F} \tag{1b}$$

In equation (1b) we use the fact that the conditional expectation is a *measurable* function with respect to $\mathcal{F}$. We do this by replacing the measure P on the left hand side with $P_{\mathcal{F}}$ to indicate that we are using the same measure, but one that is restricted to the sub σ-algebra $\mathcal{F}$.

Although it seems that f – itself – satisfies (1a), we have to be careful. Equation (1b)'s notation emphasizes that the conditional expectation function must be *measurable* with respect to the sigma algebra we are using for integration. On the right of equation (1b) we are dealing with a $\mathcal{E}$ measurable function on $\mathcal{E}$, while on the left we are dealing with a $\mathcal{F}$ measurable function. Consequently, while f seems like a nature candidate for the conditional expectation it is not necessarily $\mathcal{F}$ measurable. In fact, for any non-trivial application of conditional expectation, f is *non-measurable* with respect to the sub σ-algebra, $\mathcal{F}$. However, if f is measurable with respect to $\mathcal{F}$ then, by the definition above, it is its own conditional expectation.

The modern theory of probability – including conditional expectation – relies on measure theory which has its roots in Lebesgue measure theory – used to provide an alternative to the theory of Riemann Integration. From Lebesgue Measure Theory one is introduced to measurable and non-measurable sets/functions. The problem, in the context of Lebesgue, is that non-measurable sets/functions exist but specific examples are

† Specifically, in a discrete setting, a set Z in a σ-algebra, $\mathcal{H}$, is *minimal* in $\mathcal{H}$ if there is no non-empty, strict subset, Y, of Z with $Y \in \mathcal{H}$.

‡ That such a unique function exists is a consequence of the Radon-Nikodym theorem.

hard to come by. However, non-measurable functions appear *implicitly* in the definition of conditional expectation in that if they didn't the definition wouldn't be of any interest.

If one started out learning measure theory from the Lebesgue theory, the definition of conditional expectation might make readers somewhat uneasy as we are confronted with non-measurable sets from the very start. From this context, it is harder to get an intuitive idea of conditional expectation.

In the next section we compute the measure theoretic conditional expectation of a discrete function. In the process, we provide an intuitive idea of the measure-theoretic definition of conditional expectation which coincides with the traditional approach – except in the "singular" case which we discuss the "Redux" section of this paper.

Example Calculation of Conditional Expectation

Consider the function from the second section on an elementary dice example, $f(D_i) = i$, which is measurable in the space $(P, X, \mathcal{E})$. Let $\mathcal{F}$ be the sub σ-algebra of that section. We now compute the conditional expectation of f with respect to $\mathcal{F}$. Using (1b) we choose the *minimal* sets: $\Lambda_1 = \{D_1, D_3, D_5\}$ and $\Lambda_2 = \{D_2, D_4, D_6\}$. Since the conditional expectation function is constant on each of these sets, equation (1b) will provide a way to find the value on each set. Since the union of Λ_1 and Λ_2 constitute the entire set, X, we will know the value of the conditional expectation function on all of X; hence we will know the conditional expectation function.

Proceeding, we have

$$\int_{\Lambda_1} \mathbb{E}\left[f/\mathcal{F}\right] \, dP_{\mathcal{F}} = \int_{\Lambda_1} f \, dP \tag{2}$$

and

$$\int_{\Lambda_2} \mathbb{E}\left[f/\mathcal{F}\right] \, dP_{\mathcal{F}} = \int_{\Lambda_2} f \, dP \tag{3}$$

We know $\mathbb{E}\left[f/\mathcal{F}\right]$ is *constant* on Λ_1. We label this value as $\mathbb{E}\left[f/\mathcal{F}\right](\Lambda_1)$. From (2) we have

$$\begin{aligned}
\int_{\Lambda_1} \mathbb{E}\left[f/\mathcal{F}\right] \, dP_{\mathcal{F}} &= \int_{\Lambda_1} f \, dP \\
\mathbb{E}\left[f/\mathcal{F}\right](\Lambda_1) \int_{\Lambda_1} dP_{\mathcal{F}} &= \int_{\Lambda_1} f \, dP \\
\mathbb{E}\left[f/_{\mathcal{F}}\right](\Lambda_1) * P_{\mathcal{F}}(\Lambda_1) &= \int_{\Lambda_1} f \, dP = f(D_1) * P(D_1) + f(D_3) * P(D_3) + f(D_5) * P(D_5) && (**) \\
\mathbb{E}\left[f/\mathcal{F}\right](\Lambda_1) &= \int_{\Lambda_1} f \, dP = f(D_1) * \frac{P(D_1)}{P(\Lambda_1)} + f(D_3) * \frac{P(D_3)}{P(\Lambda_1)} + f(D_5) * \frac{P(D_5)}{P(\Lambda_1)} \\
\mathbb{E}\left[f/\mathcal{F}\right](\Lambda_1) &= \int_{\Lambda_1} f \, dP = 1 * \frac{\frac{1}{6}}{\frac{1}{2}} + 3 * \frac{\frac{1}{6}}{\frac{1}{2}} + 5 * \frac{\frac{1}{6}}{\frac{1}{2}} \\
\mathbb{E}\left[f/\mathcal{F}\right](\Lambda_1) &= \int_{\Lambda_1} f \, dP = 1 * \frac{1}{3} + 3 * \frac{1}{3} + 5 * \frac{1}{3} \\
\mathbb{E}\left[f/\mathcal{F}\right](\Lambda_1) &= 3
\end{aligned}$$

Likewise, $\mathbb{E}\left[f/\mathcal{F}\right]$ is constant on the set Λ_2. As we did above, we find the value of $\mathbb{E}\left[f/\mathcal{F}\right]$ on the set Λ_2. As with Λ_1, label $\mathbb{E}\left[f/\mathcal{F}\right](\Lambda_2)$ as the constant value of $\mathbb{E}\left[f/\mathcal{F}\right]$ on Λ_2. We have

$$
\begin{aligned}
\int_{\Lambda_2} \mathbb{E}\left[f/\mathcal{F}\right] \, dP_{\mathcal{F}} &= \int_{\Lambda_2} f \, dP \\
\mathbb{E}\left[f/\mathcal{F}\right](\Lambda_2) \int_{\Lambda_2} dP_{\mathcal{F}} &= \int_{\Lambda_2} f \, dP \\
\mathbb{E}\left[f/_{\mathcal{F}}\right](\Lambda_2) * P_{\mathcal{F}}(\Lambda_2) &= \int_{\Lambda_2} f \, dP = f(D_2) * P(D_2) + f(D_4) * P(D_4) + f(D_6) * P(D_6) \\
\mathbb{E}\left[f/\mathcal{F}\right](\Lambda_2) &= \int_{\Lambda_2} f \, dP = f(D_2) * \frac{P(D_2)}{P(\Lambda_2)} + f(D_4) * \frac{P(D_4)}{P(\Lambda_2)} + f(D_6) * \frac{P(D_6)}{P(\Lambda_2)} \\
\mathbb{E}\left[f/\mathcal{F}\right](\Lambda_2) &= \int_{\Lambda_2} f \, dP = 2 * \frac{\frac{1}{6}}{\frac{1}{2}} + 4 * \frac{\frac{1}{6}}{\frac{1}{2}} + 6 * \frac{\frac{1}{6}}{\frac{1}{2}} \\
\mathbb{E}\left[f/\mathcal{F}\right](\Lambda_2) &= \int_{\Lambda_2} f \, dP = 2 * \frac{1}{3} + 4 * \frac{1}{3} + 6 * \frac{1}{3} \\
\mathbb{E}\left[f/\mathcal{F}\right](\Lambda_2) &= 4
\end{aligned}
$$

Since a function is determined once we know what its values are on every $x \in X$, we have found the conditional expectation function, $\mathbb{E}\left[f/\mathcal{F}\right]$, as we know its values on every $x \in X$.[†]

From equation ($**$) we see that the value of the function $\mathbb{E}\left[f/\mathcal{F}\right]$ evaluated on any ‘x’ value in a minimal set, Λ, is[‡]

$$\mathbb{E}\left[f/\mathcal{F}\right](x) = \frac{\int_{\Lambda} f \, dP}{P(\Lambda)} = \int_{\Lambda} f \, dP_{\Lambda} \tag{4}$$

That is, the value of the conditional expectation on any minimal set is the *weighted average* of f over the minimal set. The weights are determined by normalizing the probability measure P over the minimal set.

More generally, discrete probability or otherwise, we may think of the σ-algebra, $\mathcal{F}$, as a coarser “mesh” than the finer “mesh”, σ-algebra $\mathcal{E}$. And we can think of the value of $\mathbb{E}\left[f/\mathcal{F}\right]$ on any “minimal” element of the mesh as the average of the function, f, over this minimal element with respect to the finer “mesh”.

From this perspective, you can think of the conditional expectation as giving the “best” representation of a function given a cruder mesh, $\mathcal{F}$, than the refined mesh, $\mathcal{E}$. Just as the “best” representation of an image at a larger pixel scale (crude mesh) would be an average over smaller scale pixels (refined mesh) of the larger pixel.

Example Redux

Let’s adjust the dice example probabilities and redo the calculation of the conditional expectation. Set the probability measure, P, as:

$$P(A) = \frac{|A \cap \{D_1, D_3, D_5\}^c|}{3} \qquad A \in \mathcal{E}$$

That is, the probability of a set, A, is the number of die in A that are not in the set $\{D_1, D_3, D_5\}$ divided by 3. With this definition, $P(\Lambda_1) = 0$ and $P(\Lambda_2) = 1$.

[†] In the “Redux” section we examine more closely what “knowing the value on every x” means.

[‡] The notation, P_Λ means that we have normalized the measure, P, so that $P_\Lambda(\Lambda) = 1$. To do this we are *assuming* that $P(\Lambda) \neq 0$.

Given P, we can find the value of $\mathbb{E}[f/\mathcal{F}]$ on the set Λ_2 as before using equation (4). However, we can't use it to find the value $\mathbb{E}[f/\mathcal{F}]$ on the set Λ_1 as $P(\Lambda_1) = 0$.

It turns out, we don't have to determine the values with any specificity on Λ_1. We can *set* the values of $\mathbb{E}[f/\mathcal{F}]$ on Λ_1 to 0; or, to any other *single* value for that matter[†]. The reason for this is that the definition of conditional expectation determines a unique element in the function space on $\mathcal{F}$. However, each element in such a space is actually an equivalence class of measurable functions over $\mathcal{F}$ who differ by a set of measure 0.

We can't directly compute what the value of a representative of $\mathbb{E}[f/\mathcal{F}]$ is on the set Λ_1, but we don't have to, since it's a set of measure 0, any value on this set will give us a function that is in the equivalence class of the conditional expectation element. Therefore, we may take $\mathbb{E}[f/\mathcal{F}]$ to be 0 on Λ_1. This gives us a representative for the conditional expectation as we have values for all $x \in X$. Through this representative we can produce the associated equivalence class of functions. Consequently, we know $\mathbb{E}[f/\mathcal{F}]$ in the function space $L_2(P, X, \mathcal{F})$.

In the case of singular sets (set of probability 0) we may revise equation (4) and write that on minimal sets, $\mathbb{E}[f/\mathcal{F}]$ is constant and satisfies:

$$\mathbb{E}[f/\mathcal{F}](x) * P(\Lambda) = \int_\Lambda f\, dP \tag{5}$$

Here, with a measure theoretic methodology, we have an elegant way to talk about conditional expectation in a singular context that doesn't exist in the traditional approach.

A Discrete Conditional Expectation Result

We generalize the discussion from the previous sections and work with functions on discrete sets with a given algebra of sets.

Definition: Given a discrete set X and an algebra of sets $\mathcal{F}$, a set A is minimal in X with respect to $\mathcal{F}$ iff $A \in \mathcal{F}$ and no proper subset of A is contained in $\mathcal{F}$.

Let X be a set with a finite number of elements. For any given σ-algebra, $\mathcal{F}$, (in this case algebra) of sets of X, let $M_\mathcal{F} = \{M_i\}_{i=1}^N$ be the set of all *distinct* minimal sets of X in $\mathcal{F}$. The set $M_\mathcal{F}$ exists as we are dealing with a finite set X with a finite power set 2^X. The set $M_\mathcal{F}$ is formed simply by examining each element in the power set of X and adding it to $M_\mathcal{F}$ if it is minimal in $\mathcal{F}$.

claim: The set $M_\mathcal{F}$ is a *partition* of the set X. Meaning:

- $\forall i, j \in 1 : N \quad M_i \cap M_j = \emptyset \quad (i \neq j)$;
- $X = \bigcup_{i=1}^{N} M_i$.
- The algebra generated by the sets in $M_\mathcal{F}$ is the collection of the unions of each element in the power set of $M_\mathcal{F}$. Further, the algebra generated is $\mathcal{F}$.

We already know that minimal sets must have the first property otherwise if there were two minimal sets whose intersection was not $\emptyset$ then the sets would either be the same or they would each have a *strict* non-trivial subset. But since each of the two minimal sets is in $\mathcal{F}$, an algebra, there intersection is also. But this is a contradiction as each are minimal sets in $\mathcal{F}$.

[†] We know that Λ_1 is minimal; so, for the conditional expectation function to be measurable it must have the same value on all elements of Λ_1.

We prove the second property by contradiction. If the union of minimal sets did not span X, then the set, $Z_0 = \{\bigcup_{i=1}^{N} M_i\}^c$, is non-empty. The set Z_0 cannot be minimal as we have already accounted for all minimal sets in our collection. Set $\mathcal{H} = 2^{Z_0} \backslash Z_0$ – the power set of Z_0 less Z_0. Note that for Z_0 to be non-minimal, it cannot consist of 1 element. Therefore, $\mathcal{H}$ is a non-empty collection of sets. If none of the non-empty sets are in $\mathcal{F}$, then Z_0 is minimal – a contradiction. Therefore, there must be a non-trivial strict subset of Z_0, Z_1, that is in $\mathcal{F}$. We can repeat this argument with Z_1 to produce, Z_2, $\ldots Z_i$ producing Z_{i+1}, etc; noting, that each time this is done the number of elements in the current "Z set" decreases by at least 1. Consequently, the process cannot proceed indefinitely. Let m be the step where this process fails; that is, there is no strict subset, Z_m, of set Z_{m-1} which is a subset of $\mathcal{F}$. But this means that Z_m is minimal by definition – contradiction. Consequently, the second property holds.

The third property follows as none of the $\{M_i\}_{i=1}^N$ can produce sets of finer granularity than themselves, they behave as if they were single elements. Therefore, the only way to produce new sets is by union, and consequently, the collection of all Elements generated is the collection of the union of each element from the power set of $\{M_i\}_{i=1}^N$. Finally, if we label $\mathcal{G}$ as the algebra generated by $M_{\mathcal{F}}$, then since its generator sets, $\{M\}_{i=1}^N$ are elements of $\mathcal{F}$, it must be the case that $\mathcal{G} \subseteq \mathcal{F}$. To show equality, suppose that there is a set $F \in \mathcal{F}$ but $F \notin \mathcal{G}$. But $F = F \cap X = F \cap \bigcup_{i=1}^{N} M_i = \bigcup_{i=1}^{N} F \cap M_i$. Since for a given i, M_i is minimal in $\mathcal{F}$, and $F \in \mathcal{F}$, the intersection must be either M_i or $\emptyset$. Consequently, F is a union of an element of the power set of $\{M\}_{i=1}^N$, which has been shown to be a member of $\mathcal{G}$. Consequently, $\mathcal{G} \equiv \mathcal{F}$.

We now assume that a function, f, is a measurable function in a probability space with values in the discrete set X; with probability measure, P; and σ-algebra, $\mathcal{E}$. We assume that $\mathcal{F}$ is a sub σ-algebra of $\mathcal{E}$ and examine the expectation of the conditional expectation, $\mathbb{E}\left[f/\mathcal{F}\right]$.[†]

$$\begin{aligned}
E[\mathbb{E}\left[f/\mathcal{F}\right]] &= \int_X \mathbb{E}\left[f/\mathcal{F}\right]\, dP_{\mathcal{F}} = \int_X f\, dP && (6)\\
&= \int_{\bigcup_{i=1}^{N} M_i} f\, dP \\
&= \sum_{i=1}^{N} \int_{M_i} f\, dP \\
&= \sum_{i=1}^{N} \int_{M_i} \mathbb{E}\left[f/\mathcal{F}\right]\, dP_{\mathcal{F}} \\
&= \sum_{i=1}^{N} \mathbb{E}\left[f/\mathcal{F}\right](M_i)\, P(M_i) && (7)
\end{aligned}$$

In equation (6) we have the following identity: $E[\mathbb{E}\left[f/\mathcal{F}\right]] = E[f]$. From equation (7) we have that the expectation of the conditional Expectation of a σ-algebra generated by minimal sets is just the weighted sum of the values of the conditional Expectation on these sets.

Finally, if g is a discrete function, the collection of all non-trivial sets of the form $g^{-1}(a) \quad a \in (-\infty, \infty)$ is a finite collection of sets, $\{M_i\}_{i=1}^N$, for some positive integer N.[‡] The collection of the union of each element

[†] Remember, the conditional expectation is not a number; it is a probability function. As such, it too has an expectation.

[‡] Each of the sets from the collection, $\{M_i\}_{i=1}^N$, corresponds to N distinct a's: $\{a_i\}_{i=1}^N$, where each "a" is a real number.

from the power set of these sets generates an algebra, $\mathcal{F}$, with each M_i ($i \in Z^+, i \in [1, N]$) being the minimal sets in this algebra. We write $\mathbb{E}\left[f/g\right]$ to mean $\mathbb{E}\left[f/\mathcal{F}\right]$. Consequently, we may write the above as:

$$\begin{aligned}
E[\mathbb{E}\left[f/g\right]] &= \int_X \mathbb{E}\left[f/g\right]\, dP = \int_X f\, dP \qquad (6') \\
&= \int_{\overset{N}{\underset{i=1}{\cup}} M_i} f\, dP \\
&= \sum_{i=1}^{N} \int_{M_i} f\, dP \\
&= \sum_{i=1}^{N} \int_{M_i} \mathbb{E}\left[f/g\right]\, dP \\
&= \sum_{i=1}^{N} \mathbb{E}\left[f/g\right](M_i)\, P(M_i) \qquad (7')
\end{aligned}$$

Here, the term, $\mathbb{E}\left[f/g\right](M_i)$ is interpreted as the value of $\mathbb{E}\left[f/\mathcal{F}\right]$ on *any* element of the set M_i – as the value of the function $\mathbb{E}\left[f/\mathcal{F}\right]$ is constant on M_i. We also drop the $dP_{\mathcal{F}}$ notation, which was only used as a crutch in the previous part of the paper. Lastly, we know from equation (5) how to compute the values of $\mathbb{E}\left[f/g\right]$ for all M_i – even when the probability of a "g" event is 0.

Equation (6') says that $E[\mathbb{E}\left[f/g\right]] = E[f]$; that is, the expectation of the conditional Expectation of f is the same as the expectation of f. Equation (7') stats that the conditional Expectation of a discrete function is the weighted average of the values of f over its level sets where the corresponding weights are probabilities of "x" values of each level set. By "x" values for level set i we mean $f^{-1}(v_i)$ where v_i is the associated f value of level set i.