A different formalism for conditional expectation

1 Intuition

Let $(\Omega, \mathcal{M}, P)$ be a sample space. Suppose there is a process occurring in this space, over time, and we have already collected data giving us part of the information about the process.

Let X be a real random variable we are trying to predict, and let O be a random variable denoting the data we have already collected. This last variable is not necessarily real; let it take values in a measure space $(V, \mathcal{E})$.

We would like to find some way to predict X from the known data O. We will, then, define a measurable function $\tilde{X} : V \to \mathbb{R}$ such that, in some way, $\tilde{X}(O) \approx X$.

Before we elaborate on what we want this 'approximately equal' sign to mean, let us first look at the specific case where X is an indicator variable of some event E. If this is the case, $\tilde{X}(O)$ would, intuitively, represent the probability that E happens given that we know the value of O. Symbolically, ignoring for the sake of argument problems of division by zero, $\tilde{X}(o) = \frac{P(E \cap O=o)}{P(O=o)}$. Hence, $\tilde{X}(o)P(O = o) = P(E \cap O = o)$, which, integrating over some possible range of values of O, would yield

$$\int_{O \in B} \tilde{X}(O) \, \mathrm{d}P = P(E \,\&\, (O \in B)). \tag{1}$$

There is no division by zero problem with this equation, and so we take it as the definition of $\tilde{X}$, with one minor observation before we fully commit to it.

Note that $P(E \,\&\, (O \in B)) = \int_{O \in B} X \, \mathrm{d}P$, which leads to a more general version of equation (1):

$$\int_{O \in B} \tilde{X}(O) \, \mathrm{d}P = \int_{O \in B} X \, \mathrm{d}P. \tag{2}$$

This last equation suffers no restriction that X is an indicator variable, and so we seek to find $\tilde{X}$ in these conditions.

2 Definition

We now proceed to find $\tilde{X}$ satisfying (2).

Define a measure P_i on $(V, \mathcal{E})$, induced by O. That is,

$$P_i(B) = P(O \in B).$$

Now, define a measure on $(V, \mathcal{E})$ as

$$\nu(B) = \int_{O \in B} X \, \mathrm{d}P.$$

Clearly, $\nu \ll P_i$, as if $P_i(B) = 0$ then

$$\begin{aligned} \nu(B) &= \int_{O \in B} X \, \mathrm{d}P \\ &= \int X \cdot \mathbb{1}_{O \in B} \, \mathrm{d}P \end{aligned}$$

And this last function is zero almost everywhere, which means the integral equals zero.

Hence, by Radon-Nikodym, **under the assumption that ν is finite**, that is, X has finite first moment, there exists some function $f : V \to \mathbb{R}$ such that, for all $B \in \mathcal{E}$,

$$\int_B f \, \mathrm{d}P_i = \int_{O \in B} X \, \mathrm{d}P.$$

Finally, we make the observation that, for any measurable function g,

$$\int g \, \mathrm{d}P_i = \int g(O) \, \mathrm{d}P,$$

and therefore

$$\int_B f \, \mathrm{d}P_i = \int_{O \in B} f(O) \, \mathrm{d}P$$

and we then conclude, finally, that for all $B \in \mathcal{E}$

$$\int_{O \in B} f(O) \, \mathrm{d}P = \int_{O \in B} X \, \mathrm{d}P.$$

Comparing with (2), we conclude f is precisely the $\tilde{X}$ we were looking for, completing the proof of existence.

We proceed to prove uniqueness.

Let $\tilde{X}$ and $\hat{X}$ be two functions satisfying (2). We will show they are equal a.e.

We know $\int_B \tilde{X} \, \mathrm{d}P_i = \int_B \hat{X} \, \mathrm{d}P_i$ for all $B \in \mathcal{E}$, which, by standard arguments, means $P_i(\tilde{X} \neq \hat{X}) = 0$. That is, $\tilde{X} = \hat{X}$ a.e. on $(V, \mathcal{E}, P_i)$. Furthermore, $P(\tilde{X}(O) \neq \hat{X}(O)) = 0$, that is, the estimate given by $\tilde{X}$ will almost always be the same as given by $\hat{X}$.

3 Estimating Estimates (Or: The Tower Law)

We have already shown that, under fixed observed data O, there is one and only one estimate for a random variable X, modulo a null set. We will now investigate what happens when there is more than one data point being observed.

Let O' be another random variable, also representing observed data, taking values in the measure space $(V', \mathcal{E}')$. We will say O' is *recoverable from* O if there is a measurable function F such that $O' = F(O)$ almost surely. In other words, O contains enough info to recover O'. This should imply that estimates built from O are, in a sense, stronger than those built from O'.

Consider the following diagram.

$$\begin{tikzcd}
\Omega \arrow[r, "O"] \arrow[rr, bend left, "O'"] \arrow[dr, "X"'] & V \arrow[r, "F"] \arrow[d, "Y"] & V' \arrow[dl, "Y'"] \\
& \mathbb{R} &
\end{tikzcd}$$

In this diagram, $Y(O)$ represents the estimate of X using data O, and $Y'(O')$ the estimate using data O'.

We assert the following: consider the r.v. $Z = Y(O)$. It is the result of estimating X. This estimate can be itself estimated by using the information O', to yield a r.v $\tilde{Z}(O')$. That is, we are using weaker information to estimate what we would estimate if we had stronger information.

This turns out to yield the same result as estimating X directly. That is:

$$\tilde{Z}(O') = Y'(O') \text{ a.s.}$$

Intuitively, to estimate an estimate of X is to estimate X.

The proof is not particularly difficult. Indeed, all we need to show is that $\tilde{Z}$ satisfies the condition defining Y'.

Fix some $B' \in \mathcal{E}'$. Let us calculate $\int_{O' \in B'} \tilde{Z}(O') \,\mathrm{d}P$.

$$\begin{aligned}
\int_{O' \in B'} \tilde{Z}(O') \,\mathrm{d}P &= \int_{O' \in B'} Z \,\mathrm{d}P \\
&= \int_{F(O) \in B'} Y(O) \,\mathrm{d}P \\
&= \int_{O \in F^{-1}(B')} Y(O) \,\mathrm{d}P \\
&= \int_{O \in F^{-1}(B')} X \,\mathrm{d}P \\
&= \int_{O' \in B'} X \,\mathrm{d}P
\end{aligned}$$

This concludes the proof.

An active ingredient in this proof is that which we refer to as the conditional expectation. Consider an r.v. X and some observed data O. The function $\widetilde{X}$ allows us to estimate X from O, but for theoretical purposes it is useful for us to inspect the r.v. $\widetilde{X}(O)$; that is, the r.v. 'the result of estimating X'. This r.v. corresponds to the usual definition of conditioning over a variable, and we denote it $E[X \mid O]$. What we have just shown is a part of the tower law: if O' is recoverable from O, then $E[E[X \mid O] \mid O'] = E[X \mid O']$.

We note that the other rest of the tower law is a triviality: Indeed, suppose we wish to estimate $E[X \mid O']$ using the data O. Well, if we know data O, then we can also recover O', and so find exactly $E[X \mid O']$.

Formally, let Y' be the estimate of X using data O'. Note that the function $Y : V \to \mathbb{R}$ defined as $Y(v) = Y'(F(v))$ satisfies the requirement (2) to be an estimate of $E[X \mid O']$, as fixed any $B \in \mathcal{E}$

$$\begin{aligned} \int_{O\in B} Y(O) &= \int_{O\in B} Y'(F(O)) \\ &= \int_{O\in B} Y'(O') \\ &= \int_{O\in B} E[X \mid O'] \end{aligned}$$

This shows that $E[E[X \mid O'] \mid O] = E[X \mid O']$. Note that the only thing we used about $E[X \mid O']$ was that it is a function of O. Hence, we conclude a version of the tower law:

Prop 1. *Suppose O' is recoverable from O.*

- *X is a function of*[1] *O iff $E[X \mid O] = X$.*
- *$E[E[X \mid O'] \mid O] = E[X \mid O'] = E[E[X \mid O] \mid O']$.*

We point out that the condition 'X is a function of O' turns out to be equivalent to 'X is measurable in the σ-algebra generated by O'.

Proof. Clearly, if $X = T(O)$ then $\{X \in B\} = \{O \in T^{-1}(B)\}$, which shows the right-implication.

The left implication is slightly tricker, but doable through usual truncation methods. Indeed, split X into its positive and negative parts, truncate it as in the construction of the Lebesgue integral, and take the limit. The only thing we need to show is that these truncations are a function of O. But if X_n is such a truncation, $X_n = \sum v_i \mathbb{1}_{v_i \leq X < v_{i+1}}$, where v_i ranges over the values X_n may take. However, $\{X \in [v_i, v_{i+1}[\}$ is, by hypothesis, of the form $\{O \in B_i\}$, and so $X_n = \sum v_i \mathbb{1}_{B_i}(O)$, which is a function of O. Taking the limit in n, we get X is a function of O, as desired. □

[1] We are sweeping a lot of 'almost everywhere's under the rug. This is supposed to mean that X is equal a.e. to something of the form $T(O)$.

We conclude this section with a trivial but useful remark.

Suppose O and O' are two pieces of partial information, each deducible from the other. That is, they both hold the same information. Then, it stands to reason that estimating X from one is the same as estimating it from the other. This is an easy consequence of the tower law, as

$$E[X \mid O] = E[E[X \mid O] \mid O'] = E[X \mid O'].$$

4 Full generality

The r.v. $E[X \mid O]$ actually turns out to be a particular case of the definition given in (2). Fixed O, consider the σ-algebra on Ω generated by O, that is, sets of the form $\{O \in B\}$ for $B \in \mathcal{E}$. Call it $\mathcal{N}$. Since O is measurable, $\mathcal{N} \subseteq \mathcal{M}$.

Consider the identity function id from the measure space $(\Omega, \mathcal{M})$ to $(\Omega, \mathcal{N})$. It is certainly measurable, and so we can apply our procedure to get an estimate for X from id. This nets us a random variable $\hat{X}$ that is $\mathcal{N}$-measurable (as is $E[X \mid O]$) and satisfies, for all $B \in \mathcal{N}$,

$$\int_B \hat{X} \,\mathrm{d}P = \int_B X \,\mathrm{d}P.$$

Note that $E[X \mid O]$ does as well, as

$$\begin{aligned}\int_B E[X \mid O] \,\mathrm{d}P &= \int_{O \in B'} \tilde{X}(O) \,\mathrm{d}P \\ &= \int_{O \in B'} X \,\mathrm{d}P \\ &= \int_B X \,\mathrm{d}P,\end{aligned}$$

where B' is the set such that $B = \{O \in B'\}$ (exists by definition of $\mathcal{N}$) and $\tilde{X}$ is the estimate for X obtained from O.

Hence, by uniqueness of $\hat{X}$, it must be equal to $E[X \mid O]$ a.e.

This proof can be adjusted to show that, if $\mathcal{N}$ is any σ-algebra contained in $\mathcal{M}$, the r.v. $E[X \mid \mathcal{N}]$ (in the usual sense) is obtained from our definition by using the identity function $\mathrm{id} : (\Omega, \mathcal{M}) \to (\Omega, \mathcal{N})$.