Conditional Expectation

Robert L. Wolpert
Institute of Statistics and Decision Sciences
Duke University, Durham, NC, USA

November 18, 2010

1 Conditioning

Frequently in probability and (especially Bayesian) statistics we wish to find the probability of some event A or the expectation of some random variable X, *conditionally* on some body of information— such as the occurance of another event B or the value of another random variable Z (or collection of them $\{Z_\alpha\}$). In elementary probability we encounter the usual formulas for conditional probabilities and expectations

$$\mathsf{P}[A \mid B] = \frac{\mathsf{P}[A \cap B]}{\mathsf{P}[B]} \qquad \mathsf{E}[X \mid Z] = \begin{cases} \frac{\int x\, f(x,Z)\, dx}{\int \ f(x,Z)\, dx} & X, Z \text{ jointly cont.} \\ \\ \frac{\sum x\, f(x,Z)}{\sum \ f(x,Z)} & X, Z \text{ discrete.} \end{cases}$$

but this notion breaks down either for distributions which are *not* jointly absolutely continuous or discrete, and also when we wish to condition on the value of infinitely-many (even uncountably-many) random variables $\{Z_\alpha\}$, as we will when we consider stochastic processes— there is no joint density function for $\{X, Z_\alpha\}$ even if each finite set has an absolutely continuous joint distribution.

Since information in probability theory is represented by σ-algebras (here $\sigma\{B\}$ or $\sigma\{Z_\alpha\}$), what we need are ways to express, interpret, and compute *conditional* probabilities of events and expectations of random variables, given σ-algebras. As a bonus, this will unify the notions of conditional probability and conditional expectation, for distributions that are discrete or continuous or neither. First, a tool to help us.

1.1 Lebesgue's Decomposition

Let μ and λ be two positive σ-finite measures on the same meaurable space $(\Omega, \mathcal{F})$. Call μ and λ *equivalent*, and write $\mu \equiv \lambda$, if they have the same null sets— so the notion of "a.e." is the same for both. More generally, we call λ *absolutely continuous* (AC) w.r.t. μ, and write $\lambda \ll \mu$, if $\mu(A) = 0$ implies $\lambda(A) = 0$, *i.e.*, if every μ-null set is also λ-null (so $\lambda \equiv \mu$ if and only if $\lambda \ll \mu$ and $\mu \ll \lambda$). We call μ and λ *mutually singular*, and write $\mu \perp \lambda$, if for some set $A \in \mathcal{F}$ we have $\mu(A^c) = 0$ and $\lambda(A) = 0$, so μ and λ are "concentrated" on disjoint sets.

For example— if $\lambda(A) = \int_A f(x)\mu(dx)$ for some non-negative function $f \in L_1(\Omega, \mathcal{F}, \mu)$ then $\lambda \ll \mu$; if $f > 0$ μ-*a.s.*, then also $\mu(A) = \int_A f(x)^{-1}\lambda(dx)$ and $\mu \equiv \lambda$. If for some other measure ν and some $f, g \in L_1(\Omega, \mathcal{F}, \nu)$ with

$$\mu(A) = \int_A f(x)\nu(dx) \qquad \lambda(A) = \int_A g(x)\nu(dx)$$

then $\mu \perp \lambda$ if $f(x)g(x) = 0$ for ν-a.e. $x \in \Omega$.

Theorem 1 (Lebesgue Decomposition) *Let μ, λ be two σ-finite measures on $(\Omega, \mathcal{F})$. Then there exist a unique pair λ_a, λ_s of σ-finite measures on $(\Omega, \mathcal{F})$ and a unique function $Y \in L_1(\Omega, \mathcal{F}, \mu)$ such that:*

$$\begin{aligned} \lambda &= \lambda_a + \lambda_s \\ \lambda_a &\ll \mu, \qquad \lambda_s \perp \mu \\ \lambda_a(A) &= \int_A Y(\omega)\mu(d\omega), \qquad A \in \mathcal{F}. \end{aligned}$$

Proof Sketch. Set

$$\mathcal{H} = \{h \in L_1(\Omega, \mathcal{F}, \mu) : \; h \ge 0, \; (\forall A \in \mathcal{F}) \; \int_A h d\mu \le \nu(A)$$

Show that $\mathcal{H}$ is closed under maxima, then find $\{h_n\}$ such that

$$\sup\left\{\int h_n d\mu : \; n \in \mathbb{N}\right\} = \sup\left\{\int h d\mu : \; h \in \mathcal{H}\right\}$$

and set $h := \sup h_n$, $Y = h\mathbf{1}_{\{h<\infty\}}$, and verify the statement of the Theorem. □

If $\mu(dx) = dx$ is Lebesgue measure on $\mathbb{R}^d$, for example, then this decomposes any probability distribution λ into an absolutely continuous part

$\lambda_a(dx) = Y(x)\,dx$ with pdf Y and a singular part $\lambda_s(dx)$ (the sum of the singular-continuous and discrete components). When $\lambda \ll \mu$ (so $\lambda_a = \lambda$ and $\lambda_s = 0$) the Radon-Nikodym derivative is often denoted $Y = \frac{d\lambda}{d\mu}$ or $\frac{\lambda(d\omega)}{\mu(d\omega)}$, and extends the idea of "density" from densities with respect to Lebesgue measure to those with respect to an arbitrary "reference" (or "base" or "dominating") measure μ. For example, the pmf $f(x) = \mathsf{P}[X = x]$ of an integer-valued random variable X may now be viewed as its pdf with respect to counting measure on $\mathbb{Z}$, so families of discrete distributions now have pdf's (if they take values in a common countable set), and random variables with mixed distributions (truncated normals, for example) have density functions with respect to a dominating measure that includes point masses where the distributions have atoms, and Lebesgue measure where they are absolutely continuous.

To further explore conditioning we apply Lebesgue's decomposition in a quite different way, with $\mu = \mathsf{P}$ a probability measure on $(\Omega, \mathcal{F}, \mathsf{P})$ and $\lambda(d\omega) = X\,d\mathsf{P}$ for some $X \in L_1$ a σ-finite measure to prove the important:

1.2 The Radon-Nikodym Theorem

Theorem 2 (Radon-Nikodym) *Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space, $X \in L_1(\Omega, \mathcal{F}, \mathsf{P})$, and $\mathcal{G} \subset \mathcal{F}$ a sub-σ-algebra. Then there exists a unique $Y \in L_1(\Omega, \mathcal{F}, \mathsf{P})$, which we will denote $Y = \mathsf{E}[X \mid \mathcal{G}]$ and call the "conditional expectation of X, given $\mathcal{G}$," that satisfies for every $G \in \mathcal{G}$:*

$$(\forall G \in \mathcal{G}) \qquad \mathsf{E}\; Y \mathbf{1}_G = \mathsf{E}\; X \mathbf{1}_G$$

Proof. First take X to be non-negative, $X \geq 0$. Define a measure λ on $\mathcal{G}$ (*not* on all of $\mathcal{F}$) by

$$\lambda(G) := \mathsf{E} X\, \mathbf{1}_G = \int_G X(\omega)\, \mathsf{P}(d\omega).$$

This is bounded (since $X \in L_1(\Omega, \mathcal{F}, \mathsf{P})$) and positive (since $X \geq 0$), so by Theorem 1 we can write $\lambda = \lambda_a + \lambda_s$ with $\lambda_a \ll \mathsf{P}$, $\lambda_s \perp \mathsf{P}$, and $\lambda_a(G) = \int_G Y\,d\mathsf{P}$ for some $Y \in L_1(\Omega, \mathcal{G}, \mathsf{P})$. But $\lambda \ll \mathsf{P}$ by construction, so $\lambda_s = 0$ and the Corollary follows.

For general X, consider separately the positive and negative parts $X_+ := \max(X, 0)$ and $X_- := \max(-X, 0)$ and set $Y := Y_+ - Y_-$. ☐

For events $A \in \mathcal{F}$ and sub-σ-algebras $\mathcal{G} \subseteq \mathcal{F}$ we denote the *contitional* *probability of A, given $\mathcal{G}$* by

$$\mathsf{P}[A \mid \mathcal{G}] = \mathsf{E}\left[\mathbf{1}_A \mid \mathcal{G}\right],$$

a $\mathcal{G}$-measurable random variable taking values in the interval $[0,1]$.

Of course X itself has the property that its integrals over events $G \in \mathcal{G}$ coincide with those of X— the point is that $Y = \mathsf{E}[X \mid \mathcal{G}]$ is a *$\mathcal{G}$-measurable approximation* to X (*i.e.*, one that depends only on the "information" encoded in $\mathcal{G}$) with this property. As we'll see below, if $\mathcal{F} \subseteq \mathcal{G}$ (or, more generally, if X is $\mathcal{G}$-measurable, so $\sigma(X) \subseteq \mathcal{G}$) then the best $\mathcal{G}$-measurable approximation is $\mathsf{E}[X \mid \mathcal{G}] = X$ itself; at the other extreme, if X is independent of $\mathcal{G}$, then one can do no better than the constant $\mathsf{E}[X \mid \mathcal{G}] \equiv \mathsf{E}X$.

1.2.1 Key Example: Countable Partitions

If $\mathcal{G} = \sigma\{\Lambda_n\}$ for a finite or countable partition $\{\Lambda_n\} \subset \mathcal{F}$ (so $\Lambda_m \cap \Lambda_n = \emptyset$ for $m \neq n$ and $\Omega = \cup \Lambda_n$), then for any $X \in L_1(\Omega, \mathcal{F}, \mathsf{P})$,

$$\mathsf{E}[X \mid \mathcal{G}] = \sum \mathbf{1}_{\Lambda_n} \mathsf{E}_{\Lambda_n}[X] = \sum \mathbf{1}_{\Lambda_n}(\omega) \frac{1}{\mathsf{P}[\Lambda_n]} \mathsf{E}[X\, \mathbf{1}_{\Lambda_n}]$$

is constant on partition elements and equal there to the P-weighted average value of X (omit from the sum any term with $\mathsf{P}[\Lambda_n] = 0$).

In particular— let $(\Omega, \mathcal{F}, \mathsf{P})$ be the unit interval with Lebesgue measure, and let $\mathcal{G}_n = \sigma\{(i/2^n, j/2^n]\}$, $0 \leq i < j \leq 2^n$. Note that $\mathcal{G}_n \subset \mathcal{G}_m$ for $n \leq m$ and that $\mathcal{F} = \bigvee \mathcal{G}_n$. Then for any $X \in L_1(\Omega, \mathcal{F}, \mathsf{P})$,

$$X_n = \mathsf{E}[X \mid \mathcal{G}_n] = 2^n \int_{i/2^n}^{(i+1)/2^n} X \, d\mathsf{P}, \qquad i/2^n < \omega \leq (i+1)/2^n.$$

This is our first example of a *margingale*, a sequence of random variables $X_n \in L_1(\Omega, \mathcal{F}, \mathsf{P})$ with the property that $X_n = \mathsf{E}[X_m \mid \mathcal{G}_n]$ for $n \leq m$; we'll see more soon. What happens as $n \to \infty$?

1.2.2 Properties:

- If $X = \mathbf{1}_A$ and if $\mathcal{G} = \sigma\{B\}$ for some $A, B \in \mathcal{F}$ with $0 < \mathsf{P}[B] < 1$,

$$\mathsf{P}[A \mid \mathcal{G}](\omega) = \mathsf{E}[\mathbf{1}_A \mid \sigma(B)](\omega) = \begin{cases} \mathsf{P}[A \cap B]/\mathsf{P}[B] & \omega \in B \\ \mathsf{P}[A \cap B^c]/\mathsf{P}[B^c] & \omega \notin B \end{cases}$$

  Thus, conditional expectation (given a σ-algebra $\mathcal{G}$) generalizes the notion of the contitional probability of one event A given another B.

- More generally, If $X \in L_1$ and if $\mathcal{G} = \sigma\{G_i\}$ for some (finite or countable) measurable partition $\{G_i\} \subset \mathcal{F}$, then

$$\mathsf{E}[X \mid \mathcal{G}](\omega) = \sum \mathbf{1}_{G_i}(\omega) \frac{1}{\mathsf{P}(G_i)} \int_{G_i} X(\omega)\, P(d\omega),$$

  the weighted average of X over the partition element that contains ω.

- If $X, Y \sim f(x, y)$ are jointly absolutely-continuous and if $\mathcal{G} = \sigma(Y)$,

$$\mathsf{E}[X \mid \sigma(Y)] = \frac{\int x f(x, Y)\, dx}{\int f(x, Y)\, dx}.$$

  Thus, conditional expectation (given a σ-algebra $\mathcal{G}$) generalizes the elementary notion of contional expectation (given an RV Y). What if X and Y are both discrete? What if just one is discrete? What if Y is a vector?

  To prove this property, first show that any event G is $\sigma(Y)$-measurable if and only if $\mathbf{1}_G = \phi(Y)$ a.s. for some Borel measurable ϕ (use a $\pi - \lambda$ argument), then extend from $\mathbf{1}_G$ to arbitrary $\sigma(Y)$-measurable random variables.

- If $X \in L_1(\Omega, \mathcal{F}, \mathsf{P})$ and if $X \perp\!\!\!\perp \mathcal{G}$ then

$$\mathsf{E}[X \mid \mathcal{G}] \equiv \mathsf{E}X.$$

  In particular, $\mathsf{E}[X \mid \{\Omega, \emptyset\}] = \mathsf{E}X$. Thus, conditional expectation (given a σ-algebra $\mathcal{G}$) generalizes the elementary notion of expectation.

- If $X \in L_1(\Omega, \mathcal{F}, \mathsf{P})$ and if $\mathcal{H} \subset \mathcal{G} \subset \mathcal{F}$, then

$$\mathsf{E}[X \mid \mathcal{H}] = \mathsf{E}\big[\ \mathsf{E}[X \mid \mathcal{G}]\ \big| \mathcal{H}\big]$$

  This is called the "tower" (or sometimes "smoothing") property of conditional expectation. It's especially useful when we have entire nested families (called *filtrations*) of σ-algebras $\{\mathcal{F}_n\}$ with $n < m \Rightarrow \mathcal{F}_n \subseteq \mathcal{F}_m$; for example, $\mathcal{F}_n = \sigma\{X_j : \ j \le n\}$ for a family $\{X_n\}$ of (non-necessarily-independent) random variables.

- A common use of the tower property is the calculation for $\mathcal{G}$-measurable $Y \in L_1$,

$$\mathsf{E}[XY] = \mathsf{E}\Big[\mathsf{E}[XY \mid \mathcal{G}]\Big] = \mathsf{E}\Big[\mathsf{E}[X \mid \mathcal{G}]\ Y\Big]$$

- If $X \in L_2(\Omega, \mathcal{F}, \mathsf{P})$ and $\{Y_n\} \subset L_2(\Omega, \mathcal{F}, \mathsf{P})$ then $\mathsf{E}[X \mid \sigma\{Y_n\}]$ is the orthogonal projection of X onto the linear span of $\{Y_n\}$ in the Hilbert space $L_2(\Omega, \mathcal{F}, \mathsf{P})$. Thus, conditional expectation (given a σ-algebra $\mathcal{G}$) generalizes the notion of orthogonal projection. This is the best way to compute conditional expectations in multivariate normal examples.

- Let $\{X_n\} \overset{\text{iid}}{\sim} L_1(\Omega, \mathcal{F}, \mathsf{P})$ with means $\mu = \mathsf{E}[X_n]$ and set $S_n = \sum_{j \le n} X_j$, $\mathcal{G}_n = \sigma\{X_1, ..., X_n\}$. Then for $n < m$,

$$\mathsf{E}[S_m \mid \mathcal{G}_n] = S_n + (m-n)\mu;$$

  in particular, S_n is a *martingale* if $\mu = 0$. If $\sigma^2 = \mathsf{V}X_n < \infty$, check that $(S_n - n\mu)^2 - n\sigma^2$ is also a martingale.

- All the usual integration tools and inequalities— DCT, MCT, Fatou, Jensen, Hölder and Minkowski, Markov, Chebychev, *etc.*— hold for *conditional* expecatations as well. For example,

$$\phi\big(\mathsf{E}[X \mid \mathcal{G}]\big) \le \mathsf{E}[\phi(X) \mid \mathcal{G}] \;\; a.s.$$

  for convex functions $\phi(\cdot)$ (note both sides are $\mathcal{G}$-measurable *random variables* now, not constants as in the familiar Jensen inequality, so the "almost surely" qualification is needed). If $X_n \to X$ in probability, for another example, then

$$\mathsf{E}[X_n \mid \mathcal{G}] \to \mathsf{E}[X \mid \mathcal{G}] \;\; a.s.$$

  if $|X_n| \le Y \in L_1$ is dominated in L_1 or if convergence $0 \le X_n \nearrow X$ is monotone, and also $\mathsf{E}[|X_n - X| \mid \mathcal{G}] \to 0$ *a.s.*.

1.3 Borel's Paradox

Let (X, Y) be the longitude, $0 \le X < 2\pi))$ and latitude, $-\pi/2 \le Y \le \pi/2$, of a point drawn uniformly from a sphere $\mathcal{S}$ (perhaps the globe). What is its *conditional* distribution of (X, Y), given that it lies on a great sircle $\mathcal{C}$? This famously ill-posed question helps motivate a careful consideration of conditioning. If the "great circle" is the equator $Y = 0$, the answer is the (perhaps expected) uniform distribution, with latitude $X \sim \mathsf{Un}\big([0, 2\pi)\big)$. But if the great circle is, say, the prime meridian $X = 0$, then the point is much more likely to be near the equator (where an interval of $Y = 0 \pm 1$ degree latitude has a large area) than near either pole (where it doesn't); in that case the conditional distribution of Y has density $f(y \mid x) = \frac{1}{2}\cos(y)\mathbf{1}_{[-\pi/2,\pi/2]}(y)$ for any $0 \le x < 2\pi$.

We simply cannot meaningfully condition on the null event that (X,Y) lies on a set of zero probability, such as a great circle. We *can* condition on events of positive probability, or on the σ-algebra generated by a random variable.

In *Radon spaces* (which include $\mathbb{R}^d$ and all complete separable metric spaces) these notions are closely related; in particular, we can always compute a version of the conditional expectation of one random-variable X given another Z as $\mathsf{E}[X \mid Z] = \phi_X(z)$ for the limit

$$\phi_X(z) = \limsup_{\epsilon \to 0} \mathsf{E}[X \mid \ \{|Z - z| < \epsilon\} \ ].$$

Let's use this to try to answer the question: What is the conditional distribution of the horizontal component X of a point drawn from the unit square, given that the point lies on the bottom edge? Let (X,Y) be the coordinates of a point drawn uniformly from the unit square and $0 < \epsilon < 1$. For $0 < x < 1$ we can compute

$$\mathsf{P}[X \le x \mid 0 \le Y \le \epsilon] = \frac{\epsilon x}{\epsilon} = x$$

and conclude (taking $\epsilon \to 0$) that the conditional *distribution* of X, given $Y = 0$, is the standard uniform, and hence the conditional expectation $\mathsf{E}[X \mid Y = 0] = 1/2$. Similarly if we let $R = Y/X$ be the ratio of Y to X, we can also compute

$$\mathsf{P}[X \le x \mid 0 \le R \le \epsilon] = \frac{\epsilon x^2/2}{\epsilon/2} = x^2,$$

so the conditional distribution of X, given $R = 0$, is $\mathsf{Be}(2,1)$, with conditional mean $\mathsf{E}[X \mid R = 0] = 2/3$. Note that both of these "events" on which we condition are the null event that (X,Y) lies on the bottom edge of the square— another example of Borel's paradox. Really these two different results were answers to different questions: one found the values of $\mathsf{P}[X \le x \mid \sigma\{Y\}]$ and $\mathsf{E}[X \mid \sigma\{Y\}]$, the other found $\mathsf{P}[X \le x \mid \sigma\{R\}]$ and $\mathsf{E}[X \mid \sigma\{R\}]$. Geometrically, what do events in $\sigma\{Y\}$ and those in $\sigma\{R\}$ look like in the square? For an arbitrary density $f(x)$ on the unit interval, can you find a random variable Z (a function of X and Y) such that $\{Z = 0\}$ is the bottom edge of the square and the conditional distribution of X given $Z = 0$ is $f(x)\,dx$? Are any conditions on $f(x)$ needed?