Notes on conditional expectations and distributions

Philip Boeken | November 20, 2023

Lp spaces and other preliminaries Throughout, let $(\Omega, \Sigma, \mathbb{P})$ be a probability space, and unless specified otherwise, let $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{S}$ be standard Borel spaces. Recall

$$\begin{aligned}
\mathcal{L}^0(\Omega; \mathcal{S}) &:= \{f : \Omega \to \mathcal{S} \text{ measurable}\} \\
\|f\|_{L^p(\Omega,\Sigma,\mathbb{P})} &:= \left(\int |f(\omega)|^p \mathrm{d}\mathbb{P}(\omega)\right)^{\frac{1}{p}} \text{ for all } p \in [1, \infty) \\
\mathcal{L}^p((\Omega, \Sigma, \mathbb{P}); \mathcal{S}) &:= \{f \in \mathcal{L}^0 : \|f\|_{L^p(\Omega,\Sigma,\mathbb{P})} < \infty\} \\
&\text{Let } f \sim_{\mathbb{P}} g \text{ be an equivalence relation that holds for } f, g \in \mathcal{L}^0 \text{ if } \mathbb{P}(f \neq g) = 0 \\
L^p((\Omega, \Sigma, \mathbb{P}); \mathcal{S}) &:= \mathcal{L}^p((\Omega, \Sigma, \mathbb{P}); \mathcal{S}) / \sim_{\mathbb{P}} \quad \text{for all } p \in \{0\} \cup [1, \infty).
\end{aligned}$$

Both $\mathcal{L}^1$ and L^1 are Banach spaces, $\mathcal{L}^2$ and L^2 are Hilbert spaces with the inner product $\langle f, g \rangle = \int fg \mathrm{d}\mathbb{P}$, and $L^q \subset L^p$ for $1 \leq p < q < \infty$ (or $p = 0$). We let $\mathcal{B}(\mathcal{X})$ denote the Borel σ-algebra of $\mathcal{X}$, and for $X \in \mathcal{L}^0(\Omega; \mathcal{X})$ we let $\sigma(X) := \{X^{-1}(B) : B \in \mathcal{B}(\mathcal{X})\}$. For $B \subseteq \Omega$ define $\sigma(B) := \{\emptyset, B, B^c, \Omega\}$, and the trace σ-algebra is defined as $\Sigma \cap B := \{S \cap B : S \in \Sigma\}$.

Conditional expectation (given a σ-algebra) Let $(X, Y, Z) \in \mathcal{L}^2((\Omega, \Sigma, \mathbb{P}); \mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$, $f \in \mathcal{L}^1(\mathcal{X} \times \mathcal{Y}, \mathbb{P}(X, Y); \mathcal{S})$, then

$$\begin{aligned}
\mathbb{E}[f(X, Y)] &:= \int f(x, y) \mathrm{d}\mathbb{P}(x, y) \\
\mathbb{E}[f(X, Y)|\sigma(Z)] &:= \operatorname*{arg\,min}_{E \in L^2((\Omega, \sigma(Z), \mathbb{P}); \mathcal{S})} \mathbb{E}[(E - f(X, Y))^2] \in L^2((\Omega, \sigma(Z), \mathbb{P}); \mathcal{S}).
\end{aligned}$$

Since L^2 is a Hilbert space, the element $\mathbb{E}[f(X, Y)|\sigma(Z)]$ exists and is unique. If $(X, Y, Z) \in \mathcal{L}^1((\Omega, \Sigma, \mathbb{P}); \mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$, one can use the L^2 definition and a density argument (L^2 dense in L^1) to define $\mathbb{E}[f(X, Y)|\sigma(Z)]$ as the unique element in $L^1((\Omega, \sigma(Z), \mathbb{P}); \mathcal{S})$ with

$$\int_B \mathbb{E}[f(X, Y)|\sigma(Z)](\omega) \mathrm{d}\mathbb{P}(\omega) = \int_B f(X, Y)(\omega) \mathrm{d}\mathbb{P}(\omega) \text{ for all } B \in \sigma(Z).$$

For a proof that $\mathbb{E}[f(X, Y)|\sigma(Z)]$ is well-defined, see Kallenberg (2002), Theorem 5.1.

Conditional expectation (given a random variable) Let $\mathcal{S}$ be standard Borel, $(X, Y, Z) \in \mathcal{L}^1((\Omega, \Sigma, \mathbb{P}); \mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$ and $f \in \mathcal{L}^1(\mathcal{X} \times \mathcal{Y}, \mathbb{P}(X, Y); \mathcal{S})$ be given. Considering the equivalence class $\mathbb{E}[f(X, Y)|\sigma(Z)] \in L^1((\Omega, \Sigma, \mathbb{P}); \mathcal{S})$, by the Doob-Dynkin lemma (Kallenberg, 2002, Lemma 1.13) there exists a *Doob-Dynkin representation*[1] $g \in \mathcal{L}^0(\mathcal{X}_Z; \mathcal{S})$ such that

$$\mathbb{E}[f(X, Y)|\sigma(Z)](\omega) = g(Z(\omega)) \quad \mathbb{P}\text{-a.s.}$$

and that is $\mathbb{P}(Z)$-a.e. uniquely determined; we use this function g to define

$$\mathbb{E}[f(X, Y)|Z] := (z \mapsto \mathbb{E}[f(X, Y)|Z = z]) := g \in L^1((\mathcal{X}_Z, \mathbb{P}(Z)); \mathcal{S}).$$

Note that $\mathbb{E}[f(X, Y)|Z = Z] = g(Z) = \mathbb{E}[f(X, Y)|\sigma(Z)]$.

Conditional distribution Let $(X, Y, Z) \in \mathcal{L}^1((\Omega, \Sigma, \mathbb{P}); \mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$. For $A \in \mathcal{B}(\mathcal{X}), B \in \mathcal{B}(\mathcal{Y})$, define the conditional probability as a version of the conditional expectation

$$\mathbb{P}(X \in A, Y \in B|Z) := \mathbb{E}[\mathbb{1}_A(X)\mathbb{1}_B(Y)|Z] \in \mathcal{L}^1((\mathcal{X}_Z, \mathbb{P}(Z)); [0, 1]).$$

For arbitrary measurable spaces $\mathcal{X} \times \mathcal{Y}$ it is not necessarily the case that for $\mathbb{P}(Z)$ almost all $z \in \mathcal{X}_Z$, the map $(A, B) \mapsto \mathbb{P}(X \in A, Y \in B|Z = z)$ is σ-additive, hence a probability measure. A well known result (see e.g. Kallenberg (2002), Theorem 5.3) is that if $\mathcal{X} \times \mathcal{Y}$ is standard Borel, then the map

$$((A, B), z) \mapsto \mathbb{P}(X \in A, Y \in B|Z = z)$$

is a Markov kernel that is $\mathbb{P}(Z)$ a.e. uniquely defined, i.e.

$$\mathbb{P}(X, Y|Z) \in L^0((\mathcal{X}_Z, \mathbb{P}(Z)); \mathcal{P}(\mathcal{X} \times \mathcal{Y})).$$

This is referred to as the *(regular) conditional distribution*. One refers to $\mathbb{P}(X, Y|Z) \in \mathcal{L}^0(\mathcal{X}_Z; \mathcal{P}(\mathcal{X} \times \mathcal{Y}))$ as a *version* of the conditional distribution.

[1] non-standard terminology

Disintegration Let $(X,Y) \in \mathcal{L}^1((\Omega,\Sigma,\mathbb{P});\mathcal{X}\times\mathcal{Y})$, then the conditional distribution $\mathbb{P}(Y|X) \in L^0((\mathcal{X},\mathbb{P}(X));\mathcal{P}(\mathcal{X}_Y))$ can be used to *disintegrate* the product measure $\mathbb{P}(X,Y)$ (see e.g. Kallenberg (2002), Theorem 5.4), i.e.

$$\mathbb{P}(X \in A, Y \in B) = \int_A \mathbb{P}(Y \in B|X = x)\mathrm{d}\mathbb{P}(x) \quad \text{for all } A \in \mathcal{B}(\mathcal{X}), B \in \mathcal{B}(\mathcal{Y}). \tag{1}$$

Even though the Markov kernel $\mathbb{P}(Y|X)$ is a equivalence class with respect to $\mathbb{P}(X)$, the equality holds without any almost-surely conditions. If we let $f \in \mathcal{L}^1(\mathcal{X}\times\mathcal{Y},\mathbb{P}(X,Y);\mathcal{S})$, the above can be used as follows:

$$\mathbb{E}[f(X,Y)] = \int \mathbb{E}[f(X,Y)|Y=y]\mathrm{d}\mathbb{P}(y) = \int\int f(x,y)\mathrm{d}\mathbb{P}(x|y)\mathrm{d}\mathbb{P}(y).$$

Conditional expectation (given an event) Let $(X,Y) \in \mathcal{L}^1((\Omega,\Sigma,\mathbb{P});\mathcal{X}\times\mathcal{Y})$. For $B \in \mathcal{B}(\mathcal{X})$ with $\mathbb{P}(X \in B) > 0$, define

$$\mathbb{E}[Y|X \in B] := \frac{1}{\mathbb{P}(X \in B)}\mathbb{E}[\mathbb{1}_{\{X\in B\}}Y] = \frac{1}{\mathbb{P}(X \in B)}\int_B \mathbb{E}[Y|X=x]\mathrm{d}\mathbb{P}(x), \tag{2}$$

where $\mathbb{E}[Y|X=x]$ is the Doob-Dynkin representation. Note that if $\mathbb{P}(X=x) > 0$ we indeed get $\mathbb{E}[Y|X\in\{x\}] = \mathbb{E}[Y|X=x]$. This relates to conditioning on a σ-algebra via

$$\mathbb{E}[Y|\sigma(\{X\in B\})] = \mathbb{E}[Y|X\in B]\mathbb{1}_{\{X\in B\}} + \mathbb{E}[Y|X\notin B]\mathbb{1}_{\{X\notin B\}} \in L^1((\Omega,\sigma(\{X\in B\}),\mathbb{P});\mathcal{Y}).$$

If $\mathbb{P}(X\in B) = 0$ we have $\mathbb{E}[Y|\sigma(X\in B)] = \mathbb{E}[Y|X\notin B]$ $\mathbb{P}$-a.s., so there is no obvious way to define a conditional expectation given an event of probability zero in terms of a conditional expectation given a σ-algebra. When $B=\{x\}$ with $\mathbb{P}(X=x)=0$ we define $\mathbb{E}[Y|X\in\{x\}] := \mathbb{E}[Y|X=x]$, with the r.h.s. being the Doob-Dynkin representation.

Conditional expecation (given a radom variable and an event) Let $(X,Y,Z) \in \mathcal{L}^1((\Omega,\Sigma,\mathbb{P});\mathcal{X}\times\mathcal{Y}\times\mathcal{Z})$. Define for $\mathbb{P}(X)$-almost all $x\in\mathcal{X}$ the following conditional expectation

$$\mathbb{E}[Y|X\in\{x\},Z] := \mathbb{E}[Y|X=x,Z] \in L^1((\mathcal{X},\mathbb{P}(Z|X=x));\mathcal{Y}),$$

where the r.h.s. $\mathbb{E}[Y|X,Z] \in L^1((\mathcal{X}\times\mathcal{Z},\mathbb{P}(X,Z));\mathcal{Y})$ is the Doob-Dynkin representation. This curries the function $\mathbb{E}[Y|X,Z]$ into $(x\mapsto\mathbb{E}[Y|X=x,Z]) \in L^1((\mathcal{X},\mathbb{P}(X));L^1((\mathcal{Z},\mathbb{P}(Z|X=x));\mathcal{Y}))$. That $\mathbb{E}[Y|X\in\{x\},Z] \in L^1((\mathcal{X},\mathbb{P}(Z|X=x));\mathcal{Y})$ follows since

$$\begin{aligned}\int \|\mathbb{E}[Y|X=x,Z=\cdot]\|_{L^1(\mathcal{X},\mathbb{P}(Z|X=x);\mathcal{Y})}\mathrm{d}\mathbb{P}(x) &= \int\int \mathbb{E}[Y|X=x,Z=z]\mathrm{d}\mathbb{P}(z|x)\mathrm{d}\mathbb{P}(x)\\ &= \int |\mathbb{E}[Y|X=x,Z=z]|\mathrm{d}\mathbb{P}(x,z) < \infty\\ &\implies \mathbb{P}(X\in\{x\in\mathcal{X}:\|\mathbb{E}[Y|X=x,Z=\cdot]\|_{L^1(\mathcal{X},\mathbb{P}(Z|X=x);\mathcal{Y})}<\infty\}) = 1,\end{aligned}$$

or in other words, $\mathbb{E}[Y|X=x,Z] \in L^1((\mathcal{X},\mathbb{P}(Z|X=x));\mathcal{Y})$ for $\mathbb{P}(X)$-almost all $x\in\mathcal{X}$.

Let $B\in\mathcal{B}(\mathcal{X})$ such that $\mathbb{P}(X\in B)>0$. To define the conditional expectation $\mathbb{E}[Y|X\in B,Z]$, we consider the random variable $\mathbb{1}_B\circ X$, consider the Doob-Dynkin representation $\mathbb{E}[Y|\mathbb{1}_B\circ X,Z]$, and define

$$\mathbb{E}[Y|X\in B,Z] := \mathbb{E}[Y|\mathbb{1}_B\circ X = 1,Z] \in L^1((\mathcal{Z},\mathbb{P}(Z|X\in B));\mathcal{Y}).$$

That this conditional expectation is $\mathbb{P}(Z|X\in B)$-a.e. uniquely defined follows from the preceding proof that $\mathbb{E}[Y|\mathbb{1}_B\circ X=1,Z] \in L^1((\mathcal{X},\mathbb{P}(Z|\mathbb{1}_B\circ X=1));\mathcal{Y})$ and that $\mathbb{P}(Z|\mathbb{1}_B\circ X=1) = \mathbb{P}(Z|X\in B)$.

For this type of conditional expectation we have an expression that is similar to equation 2: for $B\in\mathcal{B}(\mathcal{X})$ with $\mathbb{P}(X\in B)>0$, one can use a conditional version of the disintegration of equation 1 to verify that we have

$$\mathbb{P}(Y,X\in B|Z) = \mathbb{P}(Y|X\in B,Z)\mathbb{P}(X\in B|Z) \quad \mathbb{P}(Z)\text{-a.s.} \tag{3}$$

and hence also

$$\mathbb{E}[Y|X\in B,Z] = \frac{1}{\mathbb{P}(X\in B|Z)}\int_B \mathbb{E}[Y|X=x,Z]\mathrm{d}\mathbb{P}(x|Z) \quad \mathbb{P}(Z)\text{-a.s.}$$

The Borel-Kolmogorov paradox What precisely is the Borel-Kolmogorov paradox is unclear (to me), but it seems to consist of the observation that reparametrization of the state space changes conditional probabilities, and that conditional probabilities given an event of measure zero cannot be unambiguously approximated by probabilities given an event of positive probability. When uniformly sampling (X, Y) pairs on the unit disc $\{(x, y) \in [-1, 1] \times [-1, 1] : x^2 + y^2 \leq 1\}$, one can consider the conditional probability $\mathbb{P}(|Y| \geq \frac{1}{2}|X = 0) = \frac{1}{2}$. When transforming these (X, Y) points to their polar coordinates (Φ, R), we intuitively have coincidence of the conditional events $\{|Y| \geq \frac{1}{2}|X = 0\} \iff \{R \geq \frac{1}{2}|\Phi = \frac{\pi}{2}\}$, and one might be surprised by the fact that $\mathbb{P}(|Y| \geq \frac{1}{2}|X = 0) = \frac{1}{2} \neq \frac{3}{4} = \mathbb{P}(R \geq \frac{1}{2}|\Phi = \frac{\pi}{2})$.[2]

Some sources (the Wikipedia (2023) page on this paradox, for example) 'explain' this paradox by defining the conditional probabilities in terms of a limiting procedure, namely

$$\begin{aligned} \mathbb{P}(|Y| \geq \tfrac{1}{2}|X = 0) &= \lim_{\varepsilon \downarrow 0} \frac{\mathbb{P}(|Y| \geq \frac{1}{2}|X \in [-\varepsilon, \varepsilon])}{\mathbb{P}(X \in [-\varepsilon, \varepsilon])} \\ \mathbb{P}(R \geq \tfrac{1}{2}|\Phi = \tfrac{\pi}{2}) &= \lim_{\varepsilon \downarrow 0} \frac{\mathbb{P}(R \geq \frac{1}{2}|\Phi \in [\frac{\pi}{2} - \varepsilon, \frac{\pi}{2} + \varepsilon])}{\mathbb{P}(\Phi \in [\frac{\pi}{2} - \varepsilon, \frac{\pi}{2} + \varepsilon])}, \end{aligned} \tag{4}$$

where the r.h.s. probabilities can be explicitly calculated since the conditioning event has positive probability. The difference of the two l.h.s. probabilities is then explained by the difference of the limiting procedures: the conditioning sets are 'straight vertical slices' for the (X, Y) parametrization and sectors for (Φ, R).

Defining conditional distributions in terms of such limiting procedures is not straightforward. Rao (2005), section 3.2, formulates sets A, B and measure $\mathbb{P}$ such that for determining $\mathbb{P}(A|B)$ with $\mathbb{P}(B) = 0$, there are infinitely many decreasing sequences $B_1^m \supseteq B_2^m, ... \supseteq B$ with $m \in \mathbb{R}$ for which $\lim_{n\to\infty} \mathbb{P}(A|B_n^m)$ exists, but depends on m. This ambiguity is the reason that Kallenberg (2002); Rao (2005) (p.78) and others define the conditional distribution via the Doob-Dynkin representation.

There still is a relation between the limiting operations of equation (4) and conditional distributions (defined via Doob-Dynkin). The Besicovitch differentiation theorem states that for a metric space $(\mathcal{X}, d)$ and Borel measure $\mathbb{P}$ on $\mathcal{X}$ (that satisfy so called Besicovitch or Vitali covering conditions) we have $f(x) = \lim_{\varepsilon\downarrow 0} \frac{1}{\mathbb{P}(B_\varepsilon(x))} \int_{B_\varepsilon(x)} f(x')\mathrm{d}\mathbb{P}(x')$ $\mathbb{P}$-a.e. for all $f \in L^1_{loc}(\mathbb{P})$, where L^1_{loc} denotes locally integrable L^1 functions.[3] When the criteria of this theorem are met, we have the $\mathbb{P}(X)$-a.s. equality

$$\mathbb{E}[Y|X = x] = \lim_{\varepsilon \downarrow 0} \frac{1}{\mathbb{P}(X \in B_\varepsilon(x))} \int_{B_\varepsilon(x)} \mathbb{E}[Y|X = x']\mathrm{d}\mathbb{P}(x') = \lim_{\varepsilon \downarrow 0} \frac{\mathbb{E}[Y|X \in B_\varepsilon(x)]}{\mathbb{P}(X \in B_\varepsilon(x))}. \tag{5}$$

Rigot (2022) for example proves (Theorem 4.3) that the Besicovitch differentiation theorem holds for any probability measure $\mathbb{P}$ on $\mathbb{R}^n$ with the Euclidean distance. See Rigot (2022) for more information on Besicovitch theorem, and Rao (2005), sections 3.4, 4.5 and 7.6 for more general information on Borel-Kolmogorov type paradoxes, differentiation of measures.

References

Kallenberg, O. (2002). *Foundations of Modern Probability.* Probability and Its Applications. Springer, New York, 2nd ed edition.

Rao, M. M. (2005). *Conditional Measures and Applications.* Number 271 in Monographs and Textbooks in Pure and Applied Mathematics. Chapman & Hall/CRC, Boca Raton, 2nd ed edition.

Rigot, S. (2022). Differentiation of Measures in Metric Spaces. In Baudoin, F., Rigot, S., Savaré, G., Shanmugalingam, N., Ambrosio, L., Franchi, B., Markina, I., and Serra Cassano, F., editors, *New Trends on Analysis and Geometry in Metric Spaces*, Lecture Notes in Mathematics, pages 93–116, Cham. Springer International Publishing.

Wikipedia (2023). Borel–Kolmogorov paradox. *Wikipedia.*

[2] It should however not be surprising that when picking Φ and R uniformly, one does not get a uniform distribution on the unit disc. To become convinced, one can compare the (X, Y) and (Φ, R) plots of the following R code: `n <- 5000; df<-data.frame(X=runif(n,-1,1),Y=runif(n,-1,1)); df<-cbind(df,data.frame(R=(df$X^2+df$Y^2)^(1/2), Phi=atan(df$Y/df$X))); df<-df[df$R<=1,]; plot(df$X,df$Y); plot(df$Phi,df$R)`.

[3] For $\mathcal{X} = \mathbb{R}$, d the Euclidean distance and $\mathbb{P}$ the Lebesgue measure, this is known as the Lebesgue differentiation theorem.