1 Informal discussion

1a Conditional expectation: probabilistic intuition

Let X, Y and $Z = f(X, Y)$ be random variables, μ the joint distribution of X, Y, ν the marginal distribution of X, and μ_x the conditional distribution of Y given $X = x$. The conditional expectation is

$$\mathbb{E}\big(Z \,\big|\, X\big) = g(X)\,,$$

where the regression function g,

$$g(x) = \mathbb{E}\big(f(X,Y) \,\big|\, X = x\big) = \int f(x,y)\,\mu_x(\mathrm{d}y)\,,$$

is optimal in the following sense:

$$\min_g \mathbb{E}\,|g(X) - f(X,Y)|^2\,.$$

It is easy to see that $\mathbb{E}\,f(X,Y) = \mathbb{E}\,g(X)$, which is the formula of total (or iterated) expectation: $\mathbb{E}\,(\mathbb{E}\big(Z\,\big|\,X\big)) = \mathbb{E}\,Z$.

1b Conditional expectation: geometric intuition

$$g = Qf\,, \quad \text{where} \quad Q : L_2(\mu) \to L_2(\nu)$$

is the orthogonal projection, and $L_2(\nu)$ is embedded into $L_2(\mu)$ by $g \mapsto \big((x,y) \mapsto g(x)\big)$. Thus, $\langle f, \mathbf{1}\rangle = \langle g, \mathbf{1}\rangle$, which means $\mathbb{E}\,f(X,Y) = \mathbb{E}\,g(X)$.

1c Conditional distribution: naive idea

$$\mathbb{P}\big(Y \in A \,\big|\, X = x\big) = g_A(x)\,, \quad \text{where} \quad g_A = Q\mathbf{1}_{\mathbb{R}\times A}\,.$$

1d Conditional distribution: a difficulty

However, g_A is not a function but an equivalence class. We may choose a function, but the necessary conditions, such as additivity

$$g_{A\uplus B} = g_A + g_B\,,$$

may be violated on a negligible set (of x) that depends on A, B. (The more so for countable additivity.) The union of a continuum of negligible sets need not be negligible!

2 Conditional expectation

We have a probability measure μ on $\mathbb{R}^2$, and define ν by $\nu(A) = \mu(A \times \mathbb{R})$ for Borel sets $A \subset \mathbb{R}$.

2.1 Exercise. Prove that ν is a probability measure on $\mathbb{R}$.

We embed $L_2(\nu)$ into $L_2(\mu)$ by $g \mapsto f$, $f(x, y) = g(x)$.

2.2 Exercise. Prove that we get a linear isometric embedding, and its image is a closed linear subspace.

We introduce the orthogonal projection $Q : L_2(\mu) \to L_2(\nu)$ and define

$$\mathbb{E}\big(f(X, Y) \,\big|\, X \big) = g(X) \quad \text{where} \quad g = Qf\,. \tag{2.3}$$

2.4 Exercise. Prove that (2.3) conforms with the elementary definition

$$\mathbb{E}\big(f(X, Y) \,\big|\, X = x \big) = \sum_y f(x, y) \frac{\mathbb{P}\big(X = x, Y = y \big)}{\mathbb{P}\big(X = x \big)}$$

whenever μ is discrete.

2.5 Exercise. Prove that (2.3) conforms with the usual definition

$$\mathbb{E}\big(f(X, Y) \,\big|\, X = x \big) = \int f(x, y) \frac{f_{X,Y}(x, y)}{f_X(x)} \, \mathrm{d}y$$

whenever μ has a density (that is, is absolutely continuous).

In these two cases (discrete and absolutely continuous),

$$\mathbb{E}\big(f(X, Y) \,\big|\, X = x \big) = \int f(x, y) \, \mu_x(\mathrm{d}y)$$

for some family $(\mu_x)_{x \in \mathbb{R}}$ of probability measures on $\mathbb{R}$. If such a family exists, μ_x is called the conditional distribution of Y given $X = x$.

We may change μ_x at will on a ν-negligible set (of x). That is, $(\mu_x)_x$ should be treated as an equivalence class rather than a function. This equivalence class is unique.

2.6 Exercise. Prove that $f \ge 0$ implies $Qf \ge 0$ (pointwise inequalities) (a) assuming existence of conditional distributions, (b) in general.

Hint: (b): otherwise $(Qf)^+$ is closer to f than Qf.

2.7 Exercise. Prove that $f_1 \le f_2$ implies $Qf_1 \le Qf_2$.

2.8 Exercise. Prove that $f_n \uparrow f$ implies $Qf_n \uparrow Qf$ and $f_n \downarrow f$ implies $Qf_n \downarrow Qf$ (convergence ν-almost everywhere)
(a) assuming existence of conditional distributions, (b) in general.

Hint: (b): $f_n \uparrow f$ implies $\|f_n - f\| \to 0$.

3 Conditional distribution

See 'Results formulated', Sect. 5a.

We consider $f_y = \mathbf{1}_{\mathbb{R}\times(-\infty,y]}$ and $g_y = Qf_y$. (These are a continuum of equivalence classes.)

3.1 Exercise. Prove that

(a) $y_1 \le y_2$ implies $g_{y_1} \le g_{y_2}$;
(b) $y_n \downarrow -\infty$ implies $g_{y_n} \downarrow 0$;
(c) $y_n \uparrow +\infty$ implies $g_{y_n} \uparrow \mathbf{1}$;
(b) $y_n \downarrow y$ implies $g_{y_n} \downarrow g_y$.
(Convergence ν-almost everywhere.)

These relations hold for ν-almost all x, and the exceptional set may depend on $(y_n)_n$.

3.2 Lemma. It is possible to choose functions $G_y(\cdot)$ in the equivalence classes g_y such that the relations (a)–(d) hold for all x except for a single ν-negligible set.

We set $G(x,y) = G_y(x)$ and define μ_x by

$$\mu_x\big((-\infty,y]\big) = G(x,y)\,.$$

The equality

$$\iint f(x,y)\,\mu(\mathrm{d}x\mathrm{d}y) = \int\bigg(\int f(x,y)\,\mu_x(\mathrm{d}y)\bigg)\nu(\mathrm{d}x) \tag{3.3}$$

will be proven first for indicator functions $f = \mathbf{1}_A$, $A \subset \mathbb{R}^2$.

3.4 Exercise. Prove that (3.3) holds for $f = \mathbf{1}_{A\times(-\infty,y]}$ where $A \subset \mathbb{R}$ is a ν-measurable set and $y \in \mathbb{R}$.

By the monotone class theorem (or the π-λ theorem), (3.3) holds for $f = \mathbf{1}_A$ where $A \subset \mathbb{R}^2$ is a Borel set (or just a μ-measurable set).

Linear combinations of such functions approximate uniformly every bounded μ-measurable function.

Theorem 5a1 is thus proved.

4 Further information

4a Special cases

Two cases were discussed before: discrete and absolutely continuous. Here is a cingular case.

4.1 Exercise. Let Y be uniform on $(0,1)$ and

$$X = \begin{cases} 3Y & \text{for } Y < 1/3, \\ \frac{3}{2}(1-Y) & \text{for } Y > 1/3. \end{cases}$$

Find the conditional distribution of Y given $X = x$. Generalize it to $X = \varphi(Y)$ where φ is a piecewise smooth function. What if Y has a density f (not just constant)?

4b Generalizations

Standard (or nice) measurable spaces.

Standard (or Lebesgue-Rokhlin) probability spaces.

Disintegration of a measure on the product of two standard measurable spaces.

The case $Y(\omega) = \omega$. Regular conditional probability. Transition probability. Every sub-σ-field is $\sigma(X)$ for some X.

4c Applications

All properties of probabilities and expectations hold for *conditional* probabilities and expectations. For example, the conditional Hölder inequality:

$$\mathbb{E}\big(|XY|\,\big|\,\mathcal{E}\big) \le \big(\mathbb{E}\big(|X|^p\,\big|\,\mathcal{E}\big)\big)^{1/p}\big(\mathbb{E}\big(|Y|^q\,\big|\,\mathcal{E}\big)\big)^{1/q} \quad \text{for } \frac{1}{p} + \frac{1}{q} = 1\,.$$

Also, the conditional Monotone Convergence Theorem, etc.