spl15.tex
Lecture 15. 11.11.2010

11. Conditional expectations.

Suppose that X is a random variable, whose expectation exists (i.e. $E|X| < \infty$, or $X \in L_1$). Then EX, the expectation of X, is a scalar (a number) – non-random. The expectation operator E averages out all the randomness in X, to give its mean (a weighted average of the possible value of X, weighted according to their probability, in the discrete case). It often happens that we have *partial information* about X – for instance, we may know the value of a random variable Y which is associated with X, i.e. carries information about X. We may want to average out over the remaining randomness. This is an expectation conditional on our partial information, or more briefly a conditional expectation. This idea will be familiar already from elementary courses, in two cases:
1. *Discrete case*, based on the formula

$$P(A|B) := P(A \cap B)/P(B) \text{ if } P(B) > 0.$$

If X takes values $x_1, \cdots, x_m$ with probabilities $f_1(x_i) > 0$, Y takes values $y_1, \cdots, y_n$ with probabilities $f_2(y_j) > 0$, (X, Y) takes values (x_i, y_j) with probabilities $f(x_i, y_j) > 0$, then
(i) $f_1(x_i) = \Sigma_j f(x_i, y_j), \qquad f_2(y_j) = \Sigma_i f(x_i, y_j),$
(ii) $P(Y = y_j|X = x_i) = P(X = x_i, Y = y_j)/P(X = x_i) = f(x_i, y_j)/f_1(x_i)$

$$= f(x_i, y_j)/\Sigma_j f(x_i, y_j).$$

This is the *conditional distribution* of Y *given* $X = x_i$, written

$$f_{Y|X}(y_j|x_i) = f(x_i, y_j)/f_1(x_i) = f(x_i, y_j)/\Sigma_j f(x_i, y_j).$$

Its expectation is

$$E(Y|X = x_i) = \Sigma_j y_j f_{Y|X}(y_j|x_i) = \Sigma_j y_j f(x_i, y_j)/\Sigma_j f(x_i, y_j).$$

The problem here is that this approach only works when the events on which we condition have *positive* probability, which only happens in the *discrete* case.
2. *Density case.* If (X, Y) has density $f(x, y)$,

X has density $f_1(x) := \int_{-\infty}^{\infty} f(x, y)dy, \qquad Y$ has density $f_2(y) := \int_{-\infty}^{\infty} f(x, y)dx.$

We *define* the *conditional density* of Y *given* $X = x$ by the continuous analogue of the discrete formula above:

$$f_{Y|X}(y|x) := f(x,y)/f_1(x) = f(x,y)/\int_{-\infty}^{\infty} f(x,y)dy.$$

Its expectation is

$$E(Y|X=x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy = \int_{-\infty}^{\infty} y f(x,y)dy/\int_{-\infty}^{\infty} f(x,y)dy.$$

Example: Bivariate normal distribution, $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$.

$$E(Y|X=x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1),$$

the familiar *regression line* of statistics (linear model). See e,g.
N. H. BINGHAM and John M. FRY: *Regression: Linear Models in Statistics*. Springer Undergraduate Mathematics Series (SUMS), 2010.

The problem here is that joint densities need not exist – do not exist, in general.

One of the great contributions of Kolmogorov's classic book of 1933 was the realization that measure theory – specifically, the Radon-Nikodym theorem – provides a way to treat conditioning in general, without making assumptions that we are in one of the two cases – discrete case and density case – above.

Recall that the probability triple is $(\Omega, \mathcal{A}, P)$. Suppose that $\mathcal{B}$ is a sub-σ-field of $\mathcal{A}$, $\mathcal{B} \subset \mathcal{A}$ (recall that a σ-field represents information; the big σ-field $\mathcal{A}$ represents 'knowing everything', the small σ-field $\mathcal{B}$ represents 'knowing something').

Suppose that Y is a non-negative random variable whose expectation exists: $EY < \infty$. The set-function

$$Q(B) := \int_B Y dP \qquad (B \in \mathcal{B})$$

is non-negative (because Y is), σ-additive – because

$$\int_B Y dP = \Sigma_n \int_{B_n} Y dP$$

if $B = \cup_n B_n$, B_n disjoint – and defined on the σ-algebra $\mathcal{B}$, so is a *measure* on $\mathcal{B}$. If $P(B) = 0$, then $Q(B) = 0$ also (the integral of anything over a

null set is zero), so $Q << P$. By the Radon-Nikodym theorem (L7), there exists a Radon-Nikodym derivative of Q with respect to P on $\mathcal{B}$, which is $\mathcal{B}$-measurable (in the RN theorem of L7, we had 'measurable', meaning '$\mathcal{A}$-measurable; here replace $\mathcal{A}$ by $\mathcal{B}$). Following Kolmogorov (1933), we call this Radon-Nikodym derivative the *conditional expectation* of Y *given* (or *conditional on*) $\mathcal{B}$, $E(Y|\mathcal{B})$: this is $\mathcal{B}$-measurable, integrable, and satisfies

$$\int_B Y dP = \int_B E(Y|\mathcal{B}) \lceil \mathcal{P} \qquad \forall \mathcal{B} \in \mathcal{B}. \tag{$*$}$$

In the general case, where Y is a random variable whose expectation exists ($E|Y| < \infty$) but which can take values of both signs, decompose Y as

$$Y = Y_+ - Y_-$$

and define $E(Y|\mathcal{B})$ by linearity as

$$E(Y|\mathcal{B}) := E(Y_+|\mathcal{B}) - E(Y_-|\mathcal{B}).$$

Suppose now that $\mathcal{B}$ is the σ-field generated by a random variable X: $\mathcal{B} = \sigma(X)$ (so $\mathcal{B}$ represents the information contained in X, or what we know when we know X). Then $E(Y|\mathcal{B}) = E(Y|\sigma(X))$, which is written more simply as $E(Y|X)$. Its defining property is

$$\int_B Y dP = \int_B E(Y|X) dP \qquad \forall B \in \sigma(X).$$

Similarly, if $\mathcal{B} = \sigma(X_1, \cdots, X_n)$ ($\mathcal{B}$ is the information in $(X_1, \cdots, X_n)$) we write
$E(Y|\sigma(X_1, \cdots, X_n)$ as $E(Y|X_1, \cdots, X_n)$:

$$\int_B Y dP = \int_B E(Y|X_1, \cdots, X_n) dP \qquad \forall \mathcal{B} \in \sigma(X_1, \cdots, X_n).$$

Note. 1. To check that something is a conditional expectation: we have to check that it integrates the right way over the right sets [i.e., as in (*)].
2. From (*): if two things integrate the same way over all sets $B \in \mathcal{B}$, they have the same conditional expectation given $\mathcal{B}$.
3. For notational convenience, we shall pass between $E(Y|\mathcal{B})$ and $E_{\mathcal{B}}Y$ at will.
4. The conditional expectation thus defined coincides with any we may have already encountered - in regression or multivariate analysis, for example.

However, this may not be immediately obvious. The conditional expectation defined above – via σ-fields and the Radon-Nikodym theorem – is rightly called by Williams ([**W**], p.84) 'the central definition of modern probability'. It may take a little getting used to. As with all important but non-obvious definitions, it proves its worth in action: see below for properties of conditional expectations, and for its use in studying stochastic processes, particularly martingales [which are defined in terms of conditional expectations].

12. Properties of conditional expectations.

1. $\mathcal{B} = \{\emptyset, \Omega\}$. Here $\mathcal{B}$ is the *smallest* possible σ-field (*any* σ-field of subsets of Ω contains $\emptyset$ and Ω), and represents 'knowing nothing'.

$$E(Y|\{\emptyset, \Omega\}) = EY.$$

Proof. We have to check (*) for $B = \emptyset$ and $B = \Omega$. For $B = \emptyset$ both sides are zero; for $B = \Omega$ both sides are EY. //

2. $\mathcal{B} = \mathcal{A}$. Here $\mathcal{B}$ is the *largest* possible σ-field, and represents 'knowing everything'.

$$E(Y|\mathcal{A}) = Y \qquad P - a.s.$$

Proof. We have to check (*) for *all* sets $B \in \mathcal{A}$. The only integrand that integrates like Y over *all* sets is Y itself, or a function agreeing with Y except on a set of measure zero.

Note. When we condition on $\mathcal{A}$ ('knowing everything'), we *know* Y (because we know everything). There is thus no uncertainty left in Y to average out, so taking the conditional expectation (averaging out remaining randomness) has no effect, and leaves Y unaltered.
3. If Y is $\mathcal{B}$-measurable, $E(Y|\mathcal{B}) = Y$ P-a.s.

Proof. Recall that Y is *always* $\mathcal{A}$-measurable (this is the definition of Y being a random variable). For $\mathcal{B} \subset \mathcal{A}$, Y may not be $\mathcal{B}$-measurable, but if it is, the proof above applies with $\mathcal{B}$ in place of $\mathcal{A}$.

Note. If Y is $\mathcal{B}$-measurable, when we are given $\mathcal{B}$ (that is, when we condition on it), we *know* Y. That makes Y effectively a constant, and when we take the expectation of a constant, we get the same constant.