spl16.tex
Lecture 16. 14.11.2010

4 (Pull-out property). If Y is $\mathcal{B}$-measurable, $E(YZ|\mathcal{B}) = YE(Z|\mathcal{B})$ P-a.s.

Proof. We need to show

$$\int_B YZdP = Y\int_B ZdP \qquad (B \in \mathcal{B}).$$

If $Y = I_{B'}$ is the indicator of a set $B' \in \mathcal{B}$, this holds, as both sides are $\int_{B\cap B'} ZdP$. By linearity, it holds for simple $\mathcal{B}$-measurable functions. It then extends to non-negative integrable $\mathcal{B}$-measurable functions by approximation as usual, and to the general case by taking positive and negative parts. //

Note. Williams calls this property 'taking out what is known'. To remember it: if Y is $\mathcal{B}$-measurable, then given $\mathcal{B}$ we know Y, so Y is effectively a constant, so can be taken out through the integration signs.

5 (Tower property). If $\mathcal{C} \subset \mathcal{B}$, $E[E(Y|\mathcal{B})\ |\mathcal{C}] = E[Y|\mathcal{C}]$ a.s.

Proof. $E_\mathcal{C}E_\mathcal{B}Y$ is $\mathcal{C}$-measurable, and for $C \in \mathcal{C} \subset \mathcal{B}$,

$$\int_C E_\mathcal{C}[E_\mathcal{B}Y]dP = \int_C E_\mathcal{B}YdP \qquad \text{(definition of } E_\mathcal{C} \text{ as } C \in \mathcal{C})$$

$$= \int_C YdP \qquad \text{(definition of } E_\mathcal{B} \text{ as } C \in \mathcal{B}).$$

So $E_\mathcal{C}[E_\mathcal{B}Y]$ satisfies the defining relation for $E_\mathcal{C}Y$. Being also $\mathcal{C}$-measurable, it *is* $E_\mathcal{C}Y$ (a.s.). //

5' (Tower property). If $\mathcal{C} \subset \mathcal{B}$, $E[E(Y|\mathcal{C})\ |\mathcal{B}] = E[Y|\mathcal{C}]$ a.s.

Proof. $E[Y|\mathcal{C}]$ is $\mathcal{C}$-measurable, so $\mathcal{B}$-measurable as $\mathcal{C} \subset \mathcal{B}$, so $E[.|\mathcal{B}]$ has no effect, by 3. //

Corollary. $E[E(Y|\mathcal{C})\ |\mathcal{C}] = E[Y|\mathcal{C}]$ a.s.
Thus the operation $E[.|\mathcal{C}]$ is linear and *idempotent* (doing it twice is the same as doing it once), so is a *projection*. So we can use what we know about projections, from Linear Algebra, Functional Analysis etc.

Note. The tower property (in either form) is also known as the *iterated conditional expectations property* or *coarse-averaging property*. When conditioning on two σ-fields, one larger (finer), one smaller (coarser), the coarser rubs out the effect of the finer, either way round.

6. *Role of independence.* If Y is independent of $\mathcal{B}$,

$$E(Y|\mathcal{B}) = EY \qquad a.s.$$

Proof. We require

$$E[Y]P(B) = E[Y]\int_B dP = \int_B Y dP \qquad (B \in \mathcal{B}).$$

If $Y = I_A$ is an indicator, I_A, I_B are independent, so

$$P(A \cap B) = E[I_{A\cap B}] = E[I_A.I_B] = E[I_A].E[I_B] = P(A)P(B),$$

by the Multiplication Theorem (L11). This gives the result for indicators; we extend to simple functions by linearity, and thence to the non-negative integrable case and the general case as usual. //

7. *Conditional Mean Formula.*

$$E[E(Y|\mathcal{B})] = EY \quad P-a.s.$$

Proof. Take $\mathcal{C} = \{\emptyset, \Omega\}$ in 5 and use 1. //

Example. Check this for the bivariate normal distribution considered above.

8. *Conditional Variance Formula.*

$$varY = E_X var(Y|X) + var_X E(Y|X).$$

Recall $varX := E[(X - EX)^2]$. Expanding the square,

$$varX = E[X^2 - 2X.(EX) + (EX)^2] = E(X^2) - 2(EX)(EX) + (EX)^2 = E(X^2) - (EX)^2.$$

Conditional variances can be defined in the same way. Recall that $E(Y|X)$ is constant when X is known ($= x$, say), so can be taken outside an expectation over X, E_X say. Then

$$var(Y|X) := E(Y^2|X) - [E(Y|X)]^2.$$

Take expectations of both sides over X:

$$E_X var(Y|X) = E_X[E(Y^2|X)] - E_X[E(Y|X)]^2.$$

Now $E_X[E(Y^2|X)] = E(Y^2)$, by the Conditional Mean Formula, so the right is, adding and subtracting $(EY)^2$,

$$\{E(Y^2) - (EY)^2\} - \{E_X[E(Y|X)]^2 - (EY)^2\}.$$

The first term is $varY$, by above. Since $E(Y|X)$ has E_X-mean EY, the second term is $var_X E(Y|X)$, the variance (over X) of the random variable $E(Y|X$ (random because X is). Combining, the result follows.
Interpretation. $varY$ = total variability in Y,
$E_X var(Y|X)$ = variability in Y not accounted for by knowledge of X,
$var_X E(Y|X)$ = variability in Y accounted for by knowledge of X.

Example: the bivariate normal.

$$Y|X = x \text{ is } N(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1), \sigma_2^2(1 - \rho^2)), \qquad varY = \sigma_2^2,$$

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

which has variance $(\rho\sigma_2/\sigma_1)^2 varX = (\rho\sigma_2/\sigma_1)^2\sigma_1^2 = \rho^2\sigma_2^2$,

$var(Y|X = x) = \sigma_2^2$ for all $x, var(Y|X) = \sigma_2^2(1-\rho^2), E_X var(Y|X) = \sigma_2^2(1-\rho^2)$.

Corollary. $E(Y|X)$ has the same mean as Y and smaller variance (if anything).

Proof. From the Conditional Mean Formula, $E[E(Y|X)] = EY$. Since $var(Y|X) \geq 0$, $E_X var(Y|X) \geq 0$, so $varE[Y|X] \leq varY$ from the Conditional Variance Formula.

This result has important applications in estimation theory. Suppose we are to estimate a parameter θ, and are considering a statistic X as a possible estimator (or basis for an estimator) of θ. We would naturally want X to contain all the information on θ contained within the entire sample. What (if anything) does this mean in precise terms? The answer lies in the concept of *sufficiency* ('data reduction') – one of the most important contributions to

statistics of the great English statistician R. A. (Sir Ronald) Fisher (1880-1962). In the language of sufficiency, the Conditional Variance Formula is seen as (essentially) the Rao-Blackwell Theorem, a key result in the area (see the index in your favourite Statistics book if you want more here).

13. Filtrations.

The Kolmogorov triples $(\Omega, \mathcal{F}, P)$, and the Kolmogorov conditional expectations $E(X|\mathcal{B})$, give us all the machinery we need to handle *static* situations involving randomness. To handle *dynamic* situations, involving randomness which unfolds with *time* – the essence of Stochastic Processes – we need further structure.

Suppose time evolves in integer steps, $t = 0, 1, 2, \cdots$ (so we start at time $t = 0$; we postpone continuous time). We suppose, for simplicity, that information is never lost (or forgotten): thus, as time increases we learn more. Recall that σ-fields represent information or knowledge. We thus need an increasing sequence of σ-fields $\{\mathcal{F}_n : n = 0, 1, 2, \cdots\}$,

$$\mathcal{F}_n \subset \mathcal{F}_{n+1} \qquad (n = 0, 1, 2, \cdots),$$

where $\mathcal{F}_n$ represents what we know at time n. As usual, we take the σ-fields to be *complete*, i.e., to contain all subsets of null sets as null sets. Thus $\mathcal{F}_0$ represents the initial information (if there is none, $\mathcal{F}_0 = \{\emptyset, \Omega\}$, the trivial σ-field). On the other hand,

$$\mathcal{F}_\infty := lim_{n\to\infty}\mathcal{F}_n$$

represents all we ever will know (the 'Doomsday σ-field'). Often, $\mathcal{F}_\infty$ will be $\mathcal{F}$, but not always; see e.g. [**W**], S15.8 for an interesting example.

Such a family $\{\mathcal{F}_n : n = 0, 1, 2, \cdots\}$ is called a *filtration*; a probability space endowed with such a filtration, $\{\Omega, \{\mathcal{F}_n\}, \mathcal{F}, \mathcal{P}\}$ is called a *filtered probability space*. (These definitions are due to P. A. MEYER (1934-2003) of Strasbourg; Meyer and the Strasbourg (and more generally, French) school of probabilists have been responsible for the 'general theory of [stochastic] processes', and for much of the progress in stochastic integration, since the 1960s). Since the filtration is so basic to the definition of a stochastic process, the more modern term for a filtered probability space is a *stochastic basis*.