Lecture 5

Math 50051, Topics in Probability Theory and Stochastic Processes

Conditioning and σ-fields:

The operation of taking expectations of rv is the formal equivalent of the heuristic notion of "forecasting". To forecast a rv one uses some information sometimes recorded in a rv Y as above, but many other times this information is a set of events, in fact a σ-field as defined in the previous lectures. Expectations calculated using such information are called conditional expectations with respect to σ-fields. We will see later that since the information utilized is in general different from one time to another, these sets of information (σ-fields) are themselves time-indexed.

Let X be a random variable (integrable) on a probability space $(\Omega, \mathcal{F}, P)$, and let $\mathcal{Y} \subset \mathcal{F}$ be a σ-filed. Then $E(X|\mathcal{Y})$ is a random variable, s.t.

1) $E(X|\mathcal{Y})$ is $\mathcal{Y}$-measurable, ie it is completely determined by the set of information $\mathcal{Y}$.

2) For any $A \in \mathcal{Y}$

$$\int_A E(X|\mathcal{Y})dP = \int_A XdP$$

Remark: $E(X|\mathcal{Y})$ exists and it is unique in the sense that if $X = Y$ almost everywhere, then $E(X|\mathcal{Y}) = E(Y|\mathcal{Y})$ almost everywhere.

Example:

1) If $\mathcal{Y} = \{\varnothing, \Omega\}$, then $E(X|\mathcal{Y}) = E(X)$

If $\mathcal{Y} = \{\varnothing, \Omega\}$, then a random variable that is $\mathcal{Y}$-measurable is constant. Indeed, the inverse image of any interval no matter how small it is or how large has to be $\varnothing$ or Ω. In particular, the inverse image of a point has to be $\varnothing$ or Ω. There is a point "a", s.t. its inverse image is Ω. Hence the random variable to her value a. Hence $E(X|\mathcal{Y}) = a$. But we also know that

$$\int_\Omega E(X|A)dP = \int_\Omega XdP = E(X)$$

$$\int_\Omega E(X|A)dP = a\int_\Omega dP = a$$

$$a = E(X)$$

2) If X is $\mathcal{Y}$-measurable then $E(X|\mathcal{Y}) = X$ a.s.

Indeed X verifies the two properties of the conditional expectation.

① X is $\mathcal{Y}$-measurable

② $\int_A E(X|\mathcal{Y})dP = \int_A XdP$

so X is the conditional expectation. Hence $E(X|\mathcal{P}(\Omega)) = X$.

3) Show that $E(1_A|B) = P(A|B)$.

Remark: From 3) we deduce that the relationship between expectation and probability is given by the rv. 1_A:
1) $P(A) = E(1_A)$; 2) $P(A|B) = E(1_A|B)$.

Remark: A particular case of conditioning on a $\sigma - field$ is conditioning on a random vector (a vector of rvs), defined in a natural way. Fro example, if X, Y, Z are discrete rvs then

$$f(x|Y = y, Z = z) = \frac{f(x, y, z)}{f_{Y,Z}(y, z)}.$$

But it is possible to have such a formula only if all of the rvs are discrete or continuous.

Remark: We saw that $E(X|Y)$ is a function of the rv Y that takes the value $E(X|Y = y)$ when $Y = y$. Since it is a rv we can compute its expectation.

Properties of conditional expectation

1) $E(aX + bY|\mathcal{Y}) = aE(X|\mathcal{Y}) + bE(Y|\mathcal{Y})$

Indeed, $E(aX + bY|\mathcal{Y}) = aE(X|\mathcal{Y}) + bE(Y|\mathcal{Y})$ is equivalent to show that for any $A \in \mathcal{Y}$,

$$\int_A E(aX + bY|\mathcal{Y})dP = \int_A aE(X|\mathcal{Y}) + bE(Y|\mathcal{Y})dP$$

$$\int_A E(aX + bY|\mathcal{Y})dP = \int_A aX + bY dP = a\int_A XdP + b\int_A YdP$$

$$\int_A aE(X|\mathcal{Y}) + bE(Y|\mathcal{Y})dP = a\int_A E(X|\mathcal{Y})dP + b\int_A E(Y|\mathcal{Y})dP = a\int_A XdP + b\int_A YdP$$

4) $E(E(X|\mathcal{Y})) = E(X)$ (property of total expectation)

Indeed

$$E(E(X|\mathcal{Y})) = \int_\Omega E(X|\mathcal{Y})dP = \int_\Omega XdP = E(X)$$

because $\Omega \in \mathcal{Y}$.

2) $E(XY|\mathcal{Y}) = XE(Y|\mathcal{Y})$ if X is $\mathcal{Y}$-measurable, or completely determined by $\mathcal{Y}$ (this is taking out what it is known).

3) $E(E(X|\mathcal{Y})|\mathcal{H}) = E(X|\mathcal{H})$ if $\mathcal{H} \subset \mathcal{Y}$ (If you have two sets of information and one is larger than the other one, conditioning on both is like conditioning on the smallest one.)

4) If $X \geq 0$ then $E(X|\mathcal{Y}) \geq 0$

5) If $\varphi : \mathbb{R} \to \mathbb{R}$ is a convex function (its graph holds water) and X is an integrable random variable on $(\Omega, \mathcal{F}, P)$, then $\varphi(X)$ is also intergrable and

$$\varphi(E(X|\mathcal{Y})) \leq E(\varphi(X)|\mathcal{Y}) \quad a.s.$$

for any $\mathcal{Y} \in \mathcal{F}$ (called Jensen inequality)

Independent We saw: two events $A, B \in \mathcal{F}$ are called independent if

$$P(A \cap B) = P(A)P(B)$$

In general, we say that n events $A_1, A_2, ..., A_n$ are independent if

$$P(A_1 \cap A_2 ... \cap A_k) = P(A_1)P(A_2)...P(A_k)$$

for any subset of k events out of the n, for any $k \leq n$.

If two events are disjoint (they do not have anything in common) and independent then:

$$P(A \cap B) = 0 \quad (disjoint)$$

$$P(A \cap B) = P(A)P(B) \quad (independent)$$

A independent of A:

$$P(A \cap A) = P(A)P(A) = P(A), \quad P(A) = 1, 0$$

Two random variables X and Y are independent if for any Borel sets $A, B \in \mathcal{B}(\mathbb{R})$, the two events

$$\{X \in A\} \ and \ \{Y \in B\}$$

are independent. We say that n random variables $X_1, ...X_n$ are independent if for any Borel sets $B_1, ..., B_n \in \mathcal{B}(\mathbb{R})$ the event $\{X_1 \in B_1\}, ..., \{X_n \in B_n\}$ are independent.

Two σ-fields $\mathcal{Y}$ and $\mathcal{H} \subset \mathcal{F}$ are independent if any two events $A \in \mathcal{Y}, B \in \mathcal{H}$ are independent. In particular two random variables X and Y are independent if the sigma fields $\sigma(X)$ and $\sigma(Y)$ are independent. Here by $\sigma(X)$ we understand the $\sigma-$field generated by X, the sigma field generated by inverse images of Borel sets, or the set of all information that we can get from X.

A random variable X and a σ-field $\mathcal{Y}$ are independent if the σ-fields $\sigma(X)$ and $\mathcal{Y}$ are independent.

Property: If X is independent of $\mathcal{Y}$ then $E(X|\mathcal{Y}) = E(X)$ (knowing $\mathcal{Y}$ does not give us any extra information on X).

Example: Suppose you are trapped in a house with 3 doors. Door 1 leads you back to the house after 1 day. Door 2 leads you back to the house after 2 days. Door 3 leads you to freedom after 3 days. On the first trial you are equally likely to pick any of the doors. If you pick door 1 or 2, then upon your return to the house, you immediately try again (until you are free).

(i) If you don't learn from your mistakes, what is the expected number of days until freedom?

(ii) Repeat part (i) with the assumption that you do learn from your mistakes.

Example Suppose $X_1, X_2, ...$ are independent identically distributed random variables with mean μ. Let S_n denote the partial sum $S_n = X_1 + ... + X_n$. Let $\mathcal{F}_n$ denote the information in $X_1, ...X_n$ $\mathcal{F}_m = \sigma(X_1, X_2, ...X_m)$, suppose $m < n$, then:

(i) $E(S_n|\mathcal{F}_m) = S_m + (n-m)\mu$

(ii) $E({S_n}^2|\mathcal{F}_m) = {S_m}^2 + (n-m)\sigma^2$

We need to move carefully in making the idea of convergence precise for stochastic processes, since random variables are functions on Ω, having distributions and moments. There are several ways in which they might be said to converge, so we begin with some simple results that will tie all these approaches together.

Markov inequality. Let X be a non-negative random variable. Then for any $a > 0$

$$P(X \geq a) \leq EX/a.$$

Chebyshov inequality. For any random variable X

$$P(|X| \geq a) \leq EX^2/a^2, \quad a > 0.$$

The event that infinitely many of the A_n occur is expressed as

$$\{A_n \; i.o.\} = \{A_n \; infinitely \; often\}$$

$$\bigcap_{n=1}^{\infty} \bigcup_{r=n}^{\infty} A_r^c.$$

Borel-Cantelli lemma. Let $(A_n; n \geq 1)$ be a collection of events, and let A be the event $\{A_n \; i.o.\}$ that infinitely many of the A_n occur. If $\sum_{n=0}^{\infty} P(A_n) < \infty$, then $P(A) = 0$.

Second Borel-Cantelli lemma. If $(A_n; n \geq 1)$ is a collection of independent events, and $\sum_{n=1}^{\infty} P(A_n) = \infty$, then $P(A_n \; i.o.) = 1$.

Recall that a sequence x_n of real numbers is said to converge to a limit x, as $n \to \infty$, if $|x_n - x| \to 0$, as $n \to \infty$. Clearly, when we consider X_n with distribution $F_n(x)$, the existence of a limit X with distribution $F(x)$ must depend on the properties of the sequences $|X_n - X|$, and $|F_n(x) - F(x)|$. We therefore define the events

$$A_n(\epsilon) = \{|X_n - X| > \epsilon\}, \quad where \; \epsilon > 0$$

Summation lemma. This gives a criterion for a type of convergence called *almost sure convergence*. It is straightforward to show that, as $n \to \infty$,

$$P(X_n \to X) = 1,$$

if and only if finitely many $A_n(\epsilon)$ occur, for any $\epsilon > 0$.

Convergence in probability. If, for any $\epsilon > 0$,

$$P(A_n(\epsilon)) = P(|X_n - X| > \epsilon) \to 0, \quad as \; n \to \infty,$$

then X_n is said to *converge in probability* to X. We may write $X_n \xrightarrow{P} X$.

It is trivial to see that almost sure convergence implies convergence in probability; formally

$$X_n \xrightarrow{a.s.} X \Longrightarrow X_n \xrightarrow{P} X.$$

Convergence in mean square. From Chebyshov's inequality we have

$$P(|X_n - X| > \epsilon) \leq E|X_n - X|^2/\epsilon^2.$$

Therefore, if we can show that $E|X_n - X|^2 \to 0$ as $n \to \infty$, it follows that $X_n \xrightarrow{P} X$. This is often a very convenient way of showing convergence in probability, and we give it a name: if $E|X_n - X|^2 \to 0$ as $n \to \infty$, then X_n is said to *converge in mean square* to X. We may write $X_n \xrightarrow{m.s.} X$.

However, even this weaker form of convergence sometimes fails to hold; in the last resort we may have to be satisfied with showing convergence of the distributions $F_n(x)$. This is a very weak form of convergence, as it does not even require the random variables X_n to be defined on a common probability space.

Convergence in distribution. If $F_n(x) \to F(x)$ at all the points x such that $F(x)$ is continuous, then X_n is said to *converge in distribution*. We may write $X_n \xrightarrow{D} X$.

Central limit theorem. If $EX_r = \mu$ and $0 < varX_r = \sigma^2 < \infty$, then, as $n \to \infty$,

$$P\big(\frac{S_n - n\mu}{(n\sigma^2)^{1/2}} \leq x\big) \to \Phi(x),$$

where $\Phi(x)$ is the standard normal distribution.

Weak law of large numbers. If $EX_r = \mu < \infty$, then for $\epsilon > 0$, as $n \to \infty$,

$$P\Big(\Big|\frac{S_n}{n} - \mu\Big| > \epsilon\Big) \to 0.$$

Strong law of large number. As $n \to \infty$,

$$\frac{S_n}{n} \xrightarrow{a.s.} \mu$$

for some finite constant μ, if and only if $E\,|X_r| < \infty$, and then $\mu = EX_1$.

The central limit theorem is the principal reason for the appearance of the normal (or 'bell-shaped') distribution in so many statistical and scientific contexts. The first version of this theorem was proved by Abraham de Moivre before 1733.