ma414l2.tex

Lecture 2. 19.1.2012.

5. Modes of convergence.

We need (at least) four modes of convergence – two strong, one intermediate, one weak. We begin with the strong modes.

We say that $X_n \to X$ *almost surely*, or a.s., if $X_n \to X$ with probability 1: $P(X_n \to X) = 1$. For $p > 0$, we write L^p for the space of random variables X with $E[|X|^p] < \infty$, and for $X \in L_p$ write

$$\|X\|_p := [E(|X|^p)]^{1/p}.$$

This is a norm, so also by a metric. For $X_n,\ X \in L_p$, we say $X_n \to X$ in L_p, or in pth mean, if $\|X_n - X\|_p \to 0$. By the Riesz-Fischer theorem (quote), L_p is complete: if $\|X_m - X_n\|_p \to 0$ as $m, n \to \infty$, then there is some $X \in L_p$ such that $X_n \to X$ in L_p.

Neither of these two strong modes of convergence implies the other.

For the intermediate mode, we say that $X_n \to X$ *in probability* if for all $\epsilon > 0$,

$$P(|X_n - X| > \epsilon) \to 0 \qquad (n \to \infty).$$

Each of a.s. convergence and convergence in L_p implies convergence in probability, but not conversely.

Convergence in probability is also given by a metric:

$$d(X, Y) := E\Big(\frac{|X - Y|}{1 + |X - Y|}\Big).$$

This metric is also complete.

Given any sequence X_n converging in pr, there exists some subsequence converging a.s. (this also is due to F. Riesz in 1912). We quote this. Likewise, any sequence X_n converging in pth mean has an a.s. convergent subsequence.

Convergence in distribution.

We turn now to the weak mode of convergence, which deals not with values of the random variables as above but with their distributions. If X_n, X are random variables with distribution functions F_n, F, we say that $X_n \to X$ *in distribution* (or *in law*),

$$X_n \to X \quad \text{in distribution,} \quad \text{or} \quad F_n \to F \quad \text{in distribution,}$$

if

$$Ef(X_n) \to Ef(X) \quad (n \to \infty) \quad \text{for all bounded continous functions } f,$$

equivalently, if

$$\int f(x)dF_n(x) \to \int f(x)dF(x) \qquad (n \to \infty)$$

for all such f. This mode of convergence is is also generated by a metric, the *Lévy metric*:

$$d(F,G) := \inf\{\epsilon > 0 : F(x-\epsilon) - \epsilon \leq G(x) \leq F(x+\epsilon) + \epsilon \text{ for all } x\}$$

(the French probabilist Paul LÉVY (1886-1971) in 1937) (it is not obvious, but it is true, that d is a metric): if F_n, F are distribution functions,

$$F_n \to F \quad \text{in distribution} \quad \Leftrightarrow \quad d(F_n, F) \to 0.$$

This is also equivalent to

$$F_n(x) \to F(x) \qquad (n \to \infty) \qquad \text{at all continuity points } x \text{ of } F.$$

(The restriction to continuity points x of F here is vital: take X_n, X as constants c_n, c with $c_n \to c$. We should clearly have $c_n \to c$ in distribution regarded as random variables; the distribution function F of c is 0 to the left of c and 1 at c and to the right; pointwise convergence takes place everywhere except c.)

We quote that the Lévy metric is complete.

Convergence in probability ('intermediate') implies convergence in distribution ('weak'). We quote this.

There is no converse, but there is a partial converse. If X_n converges in distribution to a *constant* c, then since the distribution function of the constant c is 0 to the left of c and 1 at c and to the right, it is easy to see that in fact $X_n \to c$ in probability.

6. Characteristic functions.

If X has distribution function F, the *characteristic function* (CF) of X is

$$\phi(t) := Ee^{itX} = \int_{-\infty}^{\infty} e^{itx} dF(x) \qquad (t \in \mathbf{R}).$$

This is also the *Fourier-Stieltjes transform* of F ('Fourier transform, Stieltjes integral').

The CF has a number of important properties.

1. *Existence.* The CF always exists (the integral defining it always converges). Indeed,

$$|\phi(t)| = |\int e^{itx} dF(x)| \leq \int |e^{itx}| dF(x) = \int 1 dF(x) = 1.$$

2. *Continuity.* The CF is continuous, indeed uniformly continuous:

$$|\phi(t+u) - \phi(t)| = |\int e^{itx}(e^{itu} - 1) dF(x)| \leq \int |e^{itu} - 1|.1 dF(x) \to 0$$

as $t \to 0$, by dominated convergence.
3. *Uniqueness.* The CF determines the distribution function uniquely (so taking the CF loses no information). This is a general property of Fourier transforms; we quote this.
4. *Inversion formula.* There is an inversion formula (due to Lévy, 1937) giving the distribution function in terms of the CF. We omit this, as the formula is rarely useful.
5. *Continuity theorem* (Lévy, 1937). (i) If F_n, F have CFs ϕ_n, ϕ, and $F_n \to F$ in distribution, then

$$\phi_n(t) \to \phi(t) \qquad (n \to \infty) \qquad \text{uniformly in } t \text{ on compact sets.}$$

(ii) Conversely, if $\phi_n(t) \to \phi(t)$ pointwise, and the limit function $\phi(t)$ is continuous at $t = 0$, then ϕ is the CF of a distribution function, F say, and $F_n \to F$ in distribution.
6. *Moments.* For a random variable X, the kth *moment* of X is defined by

$$\mu_k := E[X^k].$$

The first moment is the *mean* or *expectation*, $\mu = E[X]$. (We use notation such as μ_X if there are other random variables present. Context will show whether μ denotes a mean or a measure.) If X has k moments (finite), we can expand the exponential e^{itX} in the definition of the CF and get $\sum_{j=0}^{k} (it)^j . E[X^j]/j!$ or $\sum_{j=0}^{k} (it)^j \mu_j / j!$, plus an error term. Analogy with Taylor's Theorem in Real Analysis suggests that this error term should be $o(t^k)$ at $t \to 0$. This is true; we quote it: if X has k moments finite, its CF satisfies

$$\phi(t) = \sum_{j=0}^{k} (it)^j \mu_j / j! + o(t^k) \qquad (t \to 0).$$

7. Independence.

Recall from your first course in Probability that events A, B are called independent if

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

(independence corresponds to product measures). Since $P(A) = EI_A$, this says

$$E[I_A.I_B] = E[I_A].E[I_B].$$

We generalize this. A family of events is *independent* if for any finite subfamily A_k $(k = 1, \dots, n)$, the probability of the intersection of any subfamily is the product of the probabilities. A family of random variables is *independent* if, for any finite subfamily $\{X_k\}$ $(k = 1, \dots, n)$ and any x_k, the events $\{X_k \leq x_k\}$ are independent; equivalently, the events $\{X_k \in A_k\}$ are independent for all measurable A_k.

Theorem (Multiplication Theorem). If $X_1, \dots, X_n$ are independent and $g_1, \dots, g_n$ are measurable,
(i) $g_1(X_1), \dots, g_n(X_n)$ are independent;
(ii) If the g_k are bounded,

$$E[\prod_{k=1}^{n} g_k(X_k)] = \prod_{k=1}^{n} Eg_k(X_k).$$

Proof. (i)

$$P(g_k(X_k) \in A_k, k = 1, \dots, n) = P(X_k \in g_k^{-1}(A_k), k = 1, \dots, n)$$

$$= \prod_{k=1}^{n} P(X_k \in g_k^{-1}(A_k)) = \prod_{k=1}^{n} P(g_k(X_k) \in A_k),$$

proving independence of the $g_k(X_k)$.
(ii) For simple g_k, $g_k = \sum c_{k,i_k} I_{A_{k,i_k}}$,

$$E[\prod_{i=1}^{n} g_i(X_i)] = E[\prod_{k=1}^{n} \sum c_{k,i_k} I_{A_{k,i_k}}(X_k)].$$

By independence, on the RHS $E[\prod I] = E[I(\cap)] = P(\cap) = \prod P(.) = \prod E[I]$. The RHS thus factorizes, giving the result for simple g_k. The result extends to the general case by approximation. //

Then the joint distribution function is given by

$$F(x_1,\dots,x_n) = P(X_1 \le x_1,\dots,X_n \le x_n) = \prod_{i=1}^{n} P(X_i \le x_i) = \prod_{i=1}^{n} F_i(x_i),$$

where F_i is the distribution function of X_i. The F_i are called the *marginal* distribution functions:

Random variables are independent iff their joint distribution function factorizes into the product of the marginals.

Then

$$\phi(t_1,\dots,t_k) = \int \dots \int \exp\{i(t_1x_1 + \dots + t_kx_k\} dF(x_1,\dots,x_n)$$

$$= \prod_{j=1}^{n} \int \exp\{it_jx_j\} dF_j(x_j) = \prod_{j=1}^{n} \phi_j(t_j) :$$

Random variables are independent iff their joint CF factorizes into the product of the marginals.

Convolutions.

If X, Y are independent, with distribution functions F, G and CFs ϕ, ψ, the distribution of their sum $X+Y$ is called the *convolution* (German: Faltung) of their distributions. If $X+Y$ has distribution function H and CF χ,

$$\chi(t) := Ee^{it(X+Y)} = E[e^{itX}.e^{itY}] = E[e^{itX}].E[e^{itY}] = \phi(t).\psi(t),$$

by the Multiplication Theorem:

The CF of an independent sum is the product of the CFs.

So the CF turns the easy operation of adding independent random variables into the equally easy operation of multiplying CFs. By contrast, the situation for distribution functions is less simple. If X, Y, $X+Y$ have distribution functions F, G, H,

$$H(z) := P(X+Y \le z) = \int \int_{\{x+y\le z\}} dF(x)dG(y).$$

So

$$H(z) = \int F(z-y)dG(y) = \int G(z-x)dF(x);$$

we write either expression as $(F * G)(z)$. When F, G have densities f, g, H has density

$$h(x) = \int f(x-y)g(y)dy = \int g(x-y)f(y)dy.$$

In fact, if either of F, G has a density, so does $F * G$.

So by induction, if we add n independent random variables,
(i) the CFs multiply;
(ii) the distribution is a multiple convolution, involving $n-1$ integrations.
As n increases, $n-1$ integrations become intractable, so we use CFs.

Suppose now that $X_1, \dots, X_n, \dots$ are independent and identically distributed (iid) random variables, with distribution F, CF ϕ, mean μ and variance σ^2. Recall that the *variance* (variability) is a measure of randomness,

$$\sigma^2 := E[(X-EX)^2] = E[X^2 - 2EX.X + (EX)^2] = E[X^2] - 2EX.EX + [EX]^2 :$$

$$var\ X = E(X^2) - (EX)^2.$$

(We know from the definition that $var\ X \geq 0$; this also follows from the last equation by the Cauchy-Schwarz inequality.)

8. The Weak Law of Large Numbers (WLLN) and the Central Limit Theorem (CLT).

Recall that by Real Analysis,

$$(1 + \frac{x}{n})^n \to e^x \qquad (n \to \infty)$$

(this expresses compound interest, or exponential growth, as the limit of simple interest as the interest is compounded more and more often). This extends also to complex number z, and to $z_n \to z$:

$$(1 + \frac{z_n}{n})^n \to e^z \qquad (n \to \infty).$$

The next result is due to Lévy in 1925, and in more general form to the Russian probabilist A. Ya. KHINCHIN (1894-1956) in 1929 and to Kolmogorov in 1928/29.

Theorem (Weak Law of Large Numbers, WLLN). If X_i are iid with mean μ,

$$\frac{1}{n}\sum_1^n X_k \to \mu \qquad (n \to \infty) \qquad \text{in probability.}$$

Proof. If the X_k have CF $\phi(t)$, then as the mean μ exists $\phi(t) = 1 + i\mu t + o(t)$ as $t \to 0$. So $(X_1 + \ldots + X_n)/n$ has CF

$$E \exp\{it(X_1 + \ldots + X_n)/n\} = [\phi(t/n)]^n = [1 + \frac{i\mu t}{n} + o(1/n)]^n,$$

for fixed t and $n \to \infty$. By above, the RHS has limit $e^{i\mu t}$ as $n \to \infty$. But $e^{i\mu t}$ is the CF of the constant μ. So by Lévy's continuity theorem,

$$(X_1 + \ldots + X_n)/n \to \mu \qquad (n \to \infty) \qquad \text{in distribution.}$$

Since the limit μ is constant, by above this gives

$$(X_1 + \ldots + X_n)/n \to \mu \qquad (n \to \infty) \qquad \text{in probability. } //$$

As the name implies, the Weak LLN can be strengthened, to the Strong LLN (with a.s. convergence in place of convergence in probability). We turn to this later, but proceed with a refinement of the method above, in which we retain one more term in the Taylor expansion of the CF. Note first that the CF of the standard normal distribution $\Phi = N(0,1)$, with density $\phi(x)$ and distribution function $\Phi(x)$

$$\phi(x) := \frac{e^{-x^2/2}}{\sqrt{2\pi}}, \qquad \Phi(x) := \int_\infty^x \phi(u)du$$

is $e^{-t^2/2}$. The easiest way to show this is to show

$$\int_{-\infty}^{+\infty} e^{tx}.e^{-x^2/2}dx/\sqrt{2\pi} = e^{t^2/2}$$

by completing the square, and then replace t by it by analytic continuation to get, for real t,

$$\int_{-\infty}^{+\infty} e^{itx}.e^{-x^2/2}dx/\sqrt{2\pi} = e^{-t^2/2}$$

Or, one can use contour integration and Cauchy's theorem. For both methods, see e.g. Bingham and Fry [BF], p. 21.

Theorem (Central Limit Theorem, CLT). If $X_1, \dots X_n, \dots$ are iid with mean μ and variance σ^2, and $S_n := X_1 + \dots + X_n$, then

$$(S_n - n\mu)/(\sigma\sqrt{n}) \to \Phi = N(0,1) \qquad (n \to \infty) \qquad \text{in distribution.}$$

Proof. When we subtract μ from each X_k, we change the mean from μ to 0 and the second moment from μ_2 to the variance σ^2. So by the moments property of CFs, $X_k - \mu$ has CF $1 - \frac{1}{2}\sigma^2 t^2 + o(t^2)$ as $t \to 0$. So $X_1 + \dots + X_n - n\mu$ has CF

$$E \exp\{it(X_1 + \dots + X_n - n\mu)\} = [1 - \frac{1}{2}\sigma^2 t^2 + o(t^2)]^n \qquad (t \to 0).$$

Replace t by $t/(\sigma\sqrt{n})$ and let $n \to \infty$:

$$E \exp\{it(X_1 + \dots + X_n - n\mu)/(\sigma\sqrt{n})\} = [1 - \frac{1}{2}.\frac{t^2}{n} + o(1/n)]^n \to \exp\{-t^2/2\} \quad (n \to \infty),$$

by above. The left is the CF of $(S_n - n\mu)/(\sigma\sqrt{n})$; the right is the CF of $\Phi = N(0,1)$. By the continuity theorem for CFs, this gives

$$(S_n - n\mu)/(\sigma\sqrt{n}) \to \Phi = N(0,1) \qquad (n \to \infty) \qquad \text{in distribution. } //$$

The first result of this kind is the WLLN for Bernoulli trials (tossing a coin that falls heads with probability p, tails with probability $q := 1 - p$, due to Jakob BERNOULLI (1654-1705); *Ars conjectandi*, 1713, posth.) The general WLLN above, and its strengthening the SLLN below, constitute precise forms of the 'Law of Averages', known to the man in the street. The CLT for Bernoulli trials is due to Abraham de MOIVRE (1667-1754), *Doctrine of Chances* 1738 (de Moivre found the normal distribution in 1733), later extended by P. S. de LAPLACE (1749-1827), *Théorie Analytiques des Probabilités*, 1812. The general CLT is due to J. W. LINDEBERG (1876-1932) in 1922 (the name 'central limit theorem' is due to Pólya, also in 1922). The CLT is the precise form of the 'Law of Errors', known to the physicist in the street as saying 'errors are normally distributed about the mean'.

Note. 1. The CLT largely explains why the normal distribution is so ubiquitous in Statistics – basically, this is why Statistics works.

2. The CLT and the normal distribution are static. We shall need their dynamic counterparts. The stochastic process (dynamic counterpart) corresponding to the normal distribution is *Brownian motion* (Ch. II); that of the CLT is the Erdös-Kac-Donsker *invariance principle*.