pfsl10(15).tex

**Lecture 10. 3.11.2015**

*Proof of CLT.* When we subtract $\mu$ from each $X_k$, the mean changes from $\mu$ to 0 and the second moment from $\mu_2$ to the variance $\sigma^2$. 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 $X_1 + \ldots + X_n - n\mu$ has CF

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

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

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

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

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

**3. The strong law of large numbers (SLLN); the law of the iterated logarithm (LIL); the Glivenko-Cantelli and Kolmogorov-Smirnov theorems**.

The strong law of large numbers (SLLN) is due to Kolmogorov in 1933 (in the *Grundbegriffe*). We quote (proof omitted: see e.g. SP L14):

**Theorem (Strong Law of Large Numbers (SLLN)**, Kolmogorov (1933)). If $X_1, \ldots, X_n, \ldots$ are iid random variables in $L_1$ with mean $\mu$, then

$$S_n/n \to \mu \qquad (n \to \infty) \qquad a.s.$$

Conversely, if

$$S_n/n \to c \qquad (n \to \infty) \qquad a.s.$$

for some constant $c$, then the $X \in L_1$ (i.e. $E[|X_n|] < \infty$, and $E[X] = c$.

As its name implies, the SLLN is a much stronger statement than the WLLN. But as they hold under the same conditions, we may speak loosely of LLN, when we mean either.

*The law of the iterated logarithm (LIL).*

Observe that in the LLN we divide by $n$ and get convergence in probability (weak LLN) or almost surely (strong LLN); in the CLT we divide by $\sqrt{n}$ and get convergence in distribution. We might ask whether one can 'split the difference': divide by something between $\sqrt{n}$ and $n$, and get some other kind of convergence statement. We can. We quote (Khinchin, 1924):

**Theorem (Law of the iterated logarithm (LIL)**. If $X_1, \ldots, X_n, \ldots$ are iid random variables with mean $\mu$ and variance $\sigma^2$, $S_n := X_1 + \ldots + X_n$, then

$$\limsup \frac{(S_n - n\mu)}{\sigma\sqrt{2n \log\log n}} = +1 \quad a.s., \qquad \liminf \ldots = -1 \quad a.s.$$

All points in $[-1, 1]$ are limit points of subsequences, a.s., but no others.

We summarise the conclusion of LIL in symbols as:

$$(S_n - n\mu)/(\sigma\sqrt{2n \log\log n}) \to\to [-1, 1] \qquad a.s.$$

*The Skorohod representation theorem; Slutsky's theorem*

Recall (L8, 9) our use of complete metrics. A set $A$ in a metric space $S$ is *dense* if every point in $S$ is a limit of a convergent sequence of points in $A$ (motivating example: the rationals are dense in the reals). A metric space $S$ is called *separable* if it has a countable dense set (motivating example: the reals are separable, as the rationals are countable and dense). A complete separable metric space is called a *Polish space*.

*Note.* 1. The Polish spaces are the 'nice' spaces on which to work.

2. The term honours the pioneering work done by Polish mathematicians here, particularly between the two World Wars (before WWI Poland did not exist as a country; Polish mathematics never fully recovered from the devastation of WWII).

The *Skorohod representation theorem* (A. V. SKOROHOD (1930-2011) in 1956) says that, in a Polish space (such as the real line, or Euclidean space of higher dimensions – all we will need), if we have a sequence of random variables

$$X_n \to X \qquad (n \to \infty) \qquad \text{in distribution, (= weakly)},$$

then we can construct random variables $\xi_n$, $\xi$ such that

$$X_n =_d \xi_n, \qquad X =_d \xi$$

(using "$=_d$" for "equals in distribution"), and

$$\xi_n \to \xi \qquad a.s.(= \text{strongly}).$$

For proof, see e.g. Dudley [D], 11.4, Kallenberg [K], Th. 4.30 or Bogachev [B], Vol, 2, 8.5.

If $f$ is continuous, then

$$f(\xi_n) \to f(\xi) \qquad a.s.,$$

and

$$f(X_n) =_d f(\xi_n), \qquad f(X) =_d f(\xi).$$

So

$$f(X_n) \to f(X) \qquad \text{in distribution.}$$

This is an example of the *continuous mapping theorem* in the theory of weak convergence of probability measures, for which see e.g. [Bil].

Taking $X_n$ as a random 2-vector, and writing this as $(X_n, Y_n)$, one gets various results known as *Slutzky's theorem* (E. E. SLUTZKY (1880-1948) in 1925):
if $X_n \to X$ in distribution, $Y_n \to c$ constant in distribution (= in probability, as the limit is constant – recall the proof of the WLLN), then

$$X_n + Y_n \to X + c, \qquad X_n Y_n \to cX \qquad \text{in distribution,}$$

and for $c \neq 0$,

$$X_n / Y_n \to X/c \qquad \text{in distribution.}$$

See e.g. [C], 20.6. Similarly: if random variables $X_i$ tend to constant limits $c_i$, then any rational function $r = p/q$ (a rational function $r$ is a ratio of polynomials $p, q$) of the $X_i$ tends in distribution to the same rational function of the $c_i$. In particular, this holds for polynomials, e.g., powers ([C], 20.6).

The method above is powerful and flexible. It can be used to justify various approximations commonly used in Statistics, such as variance-stabilising transformations (below), and the delta-method (handout).
*Empiricals; The Glivenko-Cantelli theorem*

The first thing to note about Parametric Statistics is that the parametric model we choose will only ever be approximately right at best. We recall *Box's Dictum* (the English statistician George E. P. BOX (1919 –)): *all models are wrong – some models are useful.* For example: much of Statistics uses

a normal model in one form or other. But no real population will ever be exactly normal. And even if it were, when we sampled from it, we would destroy normality, e.g. by the need to *round* data to record it; rounded data is necessarily rational, but a normal distribution takes irrational values a.s.

So we avoid choosing a parametric model, and ask what can be done without it. We sample from an unknown population distribution $F$. One important tool is the *empirical* (distribution function) $F_n$ of the sample $X_1, \ldots, X_n$. This is the (random!) probability distribution with mass $1/n$ at each of the data points $X_i$. Writing $\delta_c$ for the *Dirac* distribution at $c$ – the probability measure with mass 1 at $c$, or distribution function of the constant $c$ –

$$F_n := \frac{1}{n} \sum_1^n \delta_{X_i}.$$

The next result is sometimes called the *Fundamental Theorem of Statistics*. It says that, in the limit, we can recover the population distribution from the sample: *the sample determines the population in the limit*. It is due to V. I. GLIVENKO (1897-1940) and F. P. CANTELLI (1906-1985), both in 1933, and is a uniform version of Kolmogorov's Strong Law of Large Numbers (SLLN, or just LLN), also of 1933.

**Theorem (Glivenko-Cantelli Theorem, 1933)**.

$$\sup_x |F_n(x) - F(x)| \to 0 \qquad (n \to \infty) \qquad a.s.$$

*Proof.* Think of obtaining a value $\leq x$ as Bernoulli trials, with parameter (= success probability) $p := P(X \leq x) = F(x)$. So by SLLN, for each fixed $x$,

$$F_n(x) \to F(x) \qquad a.s.,$$

as $F_n(x)$ is the proportion of successes. Now fix a finite partition $-\infty = x_1 < x_2 < \ldots < x_m = +\infty$. By monotonicity of $F$ and $F_n$,

$$\sup_x |F_n(x) - F(x)| \leq \max_k |F_n(x_k) - F(x_k)| + \max_k |F(x_{k+1}) - F(x_k)|$$

(check). Letting $n \to \infty$ and refining the partition indefinitely, we get

$$\limsup_n \sup_x |F_n(x) - F(x)| \leq \sup_x \Delta F(x) \qquad a.s.,$$

where $\Delta F(x)$ denotes the jump of $F$ (if any – there are at most countably many jumps!) at $x$. This proves the result when $F$ is continuous.