pfsl11(14).tex

Lecture 11. 4.11.2014

Proof of the Glivenko-Cantelli Theorem (continued).

In the general case, we use the Probability Integral Transformation (PIT, IS, I). Let $U_1, \ldots, U_n \ldots$ be iid uniforms, $U_n \sim U(0,1)$. Let $Y_n := g(U_n)$, where $g(t) := \sup\{x : F(x) < t\}$. By PIT, $Y_n \leq x$ iff $U_n \leq F(x)$, so the Y_n are iid with law F, like the X_n, so wlog take $Y_n = X_n$. Writing G_n for the empiricals of the U_n,

$$F_n = G_n(F).$$

Writing A for the range (set of values) of F,

$$\sup_x |F_n(x) - F(x)| = \sup_{t \in A} |G_n(t) - t| \leq \sup_{[0,1]} |G_n(t) - t|, \to 0 \qquad a.s.,$$

by the result (proved above) for the continuous case. //

If F is continuous, then the argument above shows that

$$\Delta_n := \sup_x |F_n(x) - F(x)|$$

is *independent* of F, in which case we may take $F = U(0,1)$, and then

$$\Delta_n = \sup_{t \in (0,1)} |F_n(t) - t|.$$

Here Δ_n is the *Kolmogorov-Smirnov (KS) statistic*, which by above is *distribution-free* if F is continuous. It turns out that there is a uniform CLT corresponding to the uniform LLN given by the Glivenko-Cantelli Theorem: $\Delta_n \to 0$ at rate $\sqrt{n}$. The limit distribution is known – it is the *Kolmogorov-Smirnov (KS) distribution* (Kolmogorov in 1933, N. V. SMIRNOV (1900-1966) in 1944)

$$1 - 2\sum_1^\infty (-)^{k+1} e^{-2k^2x^2} \qquad (x \geq 0).$$

It turns out also that, although this result is a limit theorem for *random variables*, it follows as a special case of a limit theorem for *stochastic processes*. Writing B for Brownian motion, B_0 for the Brownian bridge ($B_0(t) := B(t) - t$, $t \in [0,1]$),

$$Z_n := \sqrt{n}(G_n(t) - t) \to B_0(t), \qquad t \in [0,1]$$

(*Donsker's Theorem*: Monroe D. DONSKER (1925-1991) in 1951, originally, the *Erdös-Kac-Donsker Invariance Principle*). The relevant mathematics here is *weak convergence of probability measures* (under an appropriate topology). Thus, the KS distribution is that of the supremum of Brownian bridge. For background, see e.g. Kallenberg Ch. 14.

Higher dimensions.

In one dimension, the half-lines $(-\infty, x]$ form the obvious class of sets to use – e.g., by differencing they give us the half-open intervals $(a, b]$, and we know from Measure Theory that these suffice. In higher dimensions, obvious analogues are the half-spaces, orthants (sets of the form $\prod_{k=1}^{n}(-\infty, x_k]$), etc. – the geometry of Euclidean space is much richer in higher dimensions. We call a class of sets a *Glivenko-Cantelli class* if a uniform LLN holds for it, a *Donsker class* if a uniform CLT holds for it. For background, see e.g. [vdVW]. This book also contains a good treatment of the *delta method* (below) in this context – the *von Mises calculus* (Richard von MISES, 1883-1953), or *infinite-dimensional delta method.*

Variance-Stabilising Transformations

In exploratory data analysis (EDA), the scatter plot may suggest that the variance is not constant throughout the range of values of the predictor variable(s). But, the theory of the Linear Model *assumes* constant variance. Where this standing assumption seems to be violated, we may seek a systematic way to *stabilise* the variance – to make it constant (or roughly so), as the theory requires.

If the response variable is y, we do this by seeking a suitable function g (sufficiently smooth – say, twice continuously differentiable), and then *transforming* our data by

$$y \mapsto g(y).$$

Suppose y has mean μ:

$$Ey = \mu.$$

Taylor expand $g(y)$ about $y = \mu$:

$$g(y) = g(\mu) + (y - \mu)g'(\mu) + \frac{1}{2}(y - \mu)^2 g''(\mu) + \ldots$$

Suppose the bulk of the response values y are fairly closely bunched around the mean μ. Then, approximately, we can treat $y - \mu$ as small; then $(y - \mu)^2$ is negligible (at least to a first approximation, which is all we are attempting here). Then

$$g(y) \sim g(\mu) + (y - \mu)g'(\mu).$$

Take expectations: as $Ey = \mu$, $Eg(y) \sim g(\mu)$. So

$$g(y) - g(\mu) \sim g(y) - Eg(y) \sim g'(\mu)(y - \mu).$$

Square both sides:

$$[g(y) - g(\mu)]^2 \sim [g'(\mu)]^2 (y - \mu)^2.$$

Take expectations: as $Ey = \mu$ and $Eg(y) \sim g(\mu)$, this says

$$var(g(y)) \sim [g'(\mu)]^2 var(y).$$

Regression. So if

$$E(y_i|x_i) = \mu_i, \qquad var(y_i|x_i) = \sigma_i^2,$$

we use EDA to try to find some link between the means μ_i and the variances σ_i^2. Suppose we try $\sigma_i^2 = H(\mu_i)$, or

$$\sigma^2 = H(\mu).$$

Then by above,

$$var(g(y)) \sim [g'(\mu)]^2 \sigma^2 = [g'(\mu)]^2 H(\mu).$$

We want *constant variance*, c^2 say. So we want

$$[g'(\mu)]^2 H(\mu) = c^2, \qquad g'(\mu) = \frac{c}{\sqrt{H(\mu)}}, \qquad g(y) = c \int \frac{dy}{\sqrt{H(y)}}.$$

Note. The idea of variance-stabilising transformations (like so much else in Statistics) goes back to Fisher (R. A. (Sir Ronald) FISHER (1890-1962)). He found the density of the sample correlation coefficient r^2 in the bivariate normal distribution – a complicated function involving the population correlation coefficient ρ^2, simplifying somewhat in the case $\rho = 0$ (see e.g. [KS1], §16.27, 28). But Fisher's z-transformation of 1921 ([KS1], §16.33)

$$r = \tanh z, \qquad z = \frac{1}{2}\log(\frac{1+r}{1-r}), \qquad \rho = \tanh \zeta, \qquad \zeta = \frac{1}{2}\log(\frac{1+\rho}{1-\rho})$$

gives z approximately normal, with variance almost independent of ρ:

$$z \sim N(0, 1/(n-1)).$$

4. Infinite divisibility; self-decomposability; stability: $I \supset SD \supset S$

In the CLT, the limit distribution is normal, $N(0,1)$, CF $\exp\{-\frac{1}{2}t^2\}$. But

$$\exp\{-\frac{1}{2}t^2\} = [\exp\{-\frac{1}{2}t^2/n\}]^n \qquad (n = 1, 2, \ldots)$$

expresses the CF of the limit law $N(0,1)$ as the nth power of the CF of another probability law, $N(0,1/n)$. So $N(0,1)$ is the nth convolution of $N(0,1/n)$. We think of this as 'splitting $N(0,1)$ up into n independent parts': $N(0,1)$ is n times 'divisible'. We can do this for each n, so $N(0,1)$ is 'infinitely divisible'.

Similarly for X Poisson $P(\lambda)$: the CF is

$$E[e^{itX}] = \sum_{n=0}^{\infty} e^{-\lambda}\lambda^n.e^{itn}/n! = \exp\{-\lambda(1-e^{it})\} = [\exp\{-(\lambda/n)(1-e^{it})\}]^n,$$

so $P(\lambda)$ is the n-fold convolution of $P(\lambda/n)$, for each n. So the Poisson distributions are infinitely divisible (id).

We can extend this to the *compound Poisson distribution* $CP(\lambda, F)$, which is very important in the actuarial/insurance industry. Suppose that the number of claims is Poisson $P(\lambda)$, and that the claim sizes are iid, with distribution F and CF ϕ. Then conditional on the number of claims being n, the total claimed in the nth convolution F^{*n}, and the CF is ϕ^n. So the total X claimed has CF

$$E[e^{itX}] = \sum_{n=0}^{\infty} e^{-\lambda}\lambda^n.\phi(t)^n/n! = \exp\{-\lambda(1-\phi(t))\} = [\exp\{-(\lambda/n)(1-\phi(t))\}]^n.$$

So $CP(\lambda, F)$ is the n-fold convolution of $CP(\lambda/n, F)$ for each n, so is id.
But this holds much more generally.
Definition. We say that a random variable X, or its distribution F, is *infinitely divisible* (id) if for each $n = 1, 2, \ldots$, X has the same distribution as the sum of n independent identically distributed random variables. We write I for the class of infinitely divisible distributions.

It turns out that I is also the class of limit laws of row-sums of triangular arrays, as follows. We say that $\{x_{nk}\}$ $(k = 1, \ldots, k_n,\ n = 1, 2, \ldots)$ is a *triangular array* if for each n, the X_{nk} are independent;
we say that the array is *uniformly asymptotically negligible (uan*, more briefly *negligible*), if for all $\epsilon > 0$,

$$P(\max_{1 \le k \le k_n} |X_{nk}| > \epsilon) \to 0 \qquad (n \to \infty).$$