Alain Berlinet

How to get Central Limit Theorems for global errors of estimates

Applications of Mathematics, Vol. 44 (1999), No. 2, 81–96

Persistent URL: http://dml.cz/dmlcz/134407

HOW TO GET CENTRAL LIMIT THEOREMS FOR GLOBAL ERRORS OF ESTIMATES

Alain Berlinet, Montpellier

(Received March 1, 1998)

Abstract. The asymptotic behavior of global errors of functional estimates plays a key role in hypothesis testing and confidence interval building. Whereas for pointwise errors asymptotic normality often easily follows from standard Central Limit Theorems, global errors asymptotics involve some additional techniques such as strong approximation, martingale theory and Poissonization. We review these techniques in the framework of density estimation from independent identically distributed random variables, i.e., the context for which they were introduced. This will avoid the mathematical difficulties associated with more complex statistical situations in which these tools have proved to be useful.

Keywords: Central Limit Theorem, global errors, strong approximation, empirical processes, U-statistics, Poissonization

MSC 2000: 60F05, 62G05

1. Introduction

Since its first version about Bernoulli trials stated by de Moivre at the beginning of the eighteenth century, the Central Limit Theorem (CLT) has assumed a key role in Probability and Statistics. De Moivre's proof relied on the fact that $n!$ behaves asymptotically as $Cn^ne^{-n}\sqrt{n}$, where the constant C was proved by Stirling to be equal to $\sqrt{2\pi}$. Using this one can prove for any random variable Z_n with binomial distribution $\mathscr{B}(n,p)$, $p\in(0,1)$, $n\geqslant 1$ the de Moivre-Laplace formula

$$P(Z_n=k)=\frac{C_n(x)}{\sqrt{2\pi np(1-p)}}\exp\left(-\frac{x^2}{2}\right),\qquad 0\leqslant k\leqslant n,$$

where $x=(k-np)/\sqrt{np(1-p)}$. In the above formula $C_n(x)$ tends to 1 as n tends to infinity, uniformly on any bounded interval. This yields the first form of the CLT.

With the same notation and $-\infty \leqslant a < b \leqslant \infty$, it states that

$$\lim_{n \longrightarrow \infty} P\left(\frac{Z_n - np}{\sqrt{np(1-p)}} \in (a,b)\right) = \frac{1}{\sqrt{2\pi}} \int_a^b \exp\Big(-\frac{x^2}{2}\Big)\,\mathrm{d}x,$$

i.e. the limit distribution as n tends to infinity of $\Big((Z_n - np)/(\sqrt{np(1-p)})\Big)$ is the standard normal distribution $\mathscr{N}(0,1)$. One had to wait until the end of the nineteenth century to see a CLT for sums of random variables with arbitrary distribution. The works by Tchebychev using the moments method and those of Liapunov using characteristic functions appeared at that time. The CLT for independent random variables that is taught in basic Probability courses is due to Lindeberg (1922) for the sufficiency part and to Feller (1935) for the necessary one. It is stated as follows. Let $\{X_i, i = 1, 2, \ldots\}$ be a sequence of independent real random variables with zero mean and finite standard deviation σ_i and denote by I_A the indicator function of a set A. Let

$$S_n = \sum_{i=1}^{n} X_i \qquad \text{and} \qquad s_n = \left(\sum_{i=1}^{n} \sigma_i^2\right)^{1/2}.$$

As n tends to infinity we have

$$(S_n/s_n) \xrightarrow{\mathscr{D}} \mathscr{N}(0,1) \qquad \text{and} \qquad \max_{1 \leqslant i \leqslant n} \sigma_i/s_n \longrightarrow 0$$

if and only if

$$\text{(1)} \qquad \forall \varepsilon > 0, \qquad s_n^{-2} \sum_{i=1}^{n} E\left(X_i^2 I_{\{|X_i| > \varepsilon s_n\}}\right) \longrightarrow 0.$$

Condition (1) is called the Lindeberg condition. If the variables $\{X_i \colon\ i = 1, 2, \ldots\}$ have a finite moment of order 3, it is easily proved that

$$s_n^{-2} \sum_{i=1}^{n} E\left(X_i^2 I_{\{|X_i| > \varepsilon s_n\}}\right) \leqslant \varepsilon^{-1} s_n^{-3} \sum_{i=1}^{n} E\left(|X_i|^3\right),$$

in which case one often prefers to use the Liapunov condition

$$\text{(2)} \qquad s_n^{-3} \sum_{i=1}^{n} E\left(|X_i|^3\right) \longrightarrow 0 \quad \text{as} \quad n \longrightarrow \infty,$$

as a sufficient condition for asymptotic normality because it is easy to check.

The study of sums of random variables has generated a considerable literature and many generalizations in all possible directions of the Lindeberg-Feller result have appeared in the second half of the twentieth century. The main aspects of these developments are: extension to triangular arrays, characterization of the possible limit laws (infinite divisible laws), treatment of dependent variables, i.e. martingales, m-dependent and associated sequences, mixing processes, extension to vector spaces and functional forms. Historical background and references can be found in Heyde (1983).

For functional estimates the asymptotic normality of pointwise errors often follows from the standard CLT. However, in hypothesis testing and confidence interval building, information on global errors of estimates is needed. The aim of the paper is to expose the additional techniques (strong approximation, martingales, Poissonization) involved in the proofs of CLT for global errors. We restrict ourselves to density estimation from independent identically distributed (i.i.d.) random variables, i.e. the context for which these techniques were introduced, to avoid mathematical difficulties.

The estimates and errors under study are defined in Section 2 where the problem is set. A rich literature is available on the approximation of empirical processes. Hence a natural idea is to express the statistics under consideration as functionals of empirical processes and to exploit asymptotic results in this field. The method is presented in Section 3. Convergence to normality also holds for various statistics when the standard CLTs are not directly applicable (rank statistics, L- or M-estimates). We give in Section 4 an example of U-statistics and the application of a CLT for these statistics to L_2-errors. The main drawback of the techniques described in Sections 3 and 4 is the necessity of assuming analytical conditions on the function to be estimated at some stage of their application. These conditions can not be checked from a sample and therefore are not desirable in view of statistical applications. For this reason more probabilistic methods based on Poissonization were developed. They are presented in Section 5. Their interest relies on the independence properties of Poisson processes, which significantly simplify calculations.

2. Problem setting

Let $\mathbf{X} = \{X_i\colon\ i = 1, 2, \ldots\}$ be an infinite sequence of i.i.d. random variables with common unknown density f with respect to the Lebesgue measure λ on $\mathbb{R}$. We denote by μ the measure with density f, by F the distribution function of μ and by F_n the empirical distribution function defined from the sample $X_1, \ldots, X_n$, i.e.,

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \leqslant x\}}.$$

We can consider different types of estimates f_n of f based on $X_1, \dots, X_n$. As shown below most of these estimates are of the form

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_n(x, X_i), \tag{3}$$

where K_n is called the "generalized" kernel.

- **histogram**

Let $\mathscr{P}_n = \{A_{n,i} \colon i \in \{\dots, -1, 0, 1, \dots\}\}$ be a partition of $\mathbb{R}$ into Borel sets with positive measure. The standard histogram f_n associated with $\mathscr{P}_n$ is defined by

$$f_n(x) = \frac{\mu_n(A_n(x))}{\lambda(A_n(x))},$$

where $A_n(x)$ is the set of the partition $\mathscr{P}_n$ that contains x and μ_n is the empirical measure on Borel sets

$$\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} I_A(X_i).$$

Thus $f_n(x)$ is of the form (3) with

$$K_n(x, y) = \sum_j I_{A_{nj}}(x) I_{A_{nj}}(y) / \lambda(A_{nj}).$$

- **kernel estimate**

The standard Parzen-Rosenblatt kernel estimate on $\mathbb{R}$ is defined by

$$f_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)$$

where K is a real function on $\mathbb{R}$ integrating to one and $h_n > 0$ is the smoothing parameter. Therefore it can be expressed in the form (3) with

$$K_n(x, y) = \frac{1}{h_n} K\left(\frac{x - y}{h_n}\right).$$

- **Barron estimates**

Choose a reference density g from which the unknown density f is not too far in some sense to be specified later on. Let $\{m_n\}$, $0 < m_n < n$, $n \geqslant 2$, be a sequence of integers, set $h_n = 1/m_n$ and denote by ν the probability measure with density g. Next, introduce partitions $\mathscr{P}_n = \{A_{n,1}, A_{n,2}, \dots, A_{n,m_n}\}$, $n \geqslant 2$, of $\mathbb{R}$ such that the $A_{n,i}$'s are intervals with $\nu(A_{n,i}) = h_n$ and define

$$f_n(x) = ((1 - a_n)\mu_n(A_n(x))/h_n + a_n) g(x), \tag{4}$$

where $a_n = (1 + nh_n)^{-1}$ and $A_n(x) = A_{n,i}$ if $x \in A_{n,i}$. One can get this estimate by transforming first the data into $[0, 1]$ by the distribution function of g, then by constructing a histogram on $[0, 1]$ from a uniform partition into m_n intervals, taking the mixture of this histogram and the uniform density with weights $1 - a_n$ and a_n, respectively, and finally transforming this mixture back to the real line. This estimate was introduced by Barron (1988) and generalized by Barron, Györfi and van der Meulen (1992). It can be written in the form (3) with

$$K_n(x, y) = \sum_j \left(\frac{1 - a_n}{h_n} I_{A_{nj}}(y) + a_n \right) I_{A_{nj}}(x) g(x).$$

It is not difficult to see that many other sorts of estimates (orthogonal series, spline, etc.) are covered by (3). For such estimates it is clear that the wide range of limit theorems for sums of random variables will provide, for fixed x, conditions for the asymptotic normality of $(f_n(x) - f(x))$ or $(f_n(x) - Ef_n(x))$, suitably normalized.

Now, let $d(f, g)$ be a distance or divergence between the densities f and g and let us consider functions $\varphi[d(f_n, g_n)]$ of global errors of estimates $d(f_n, g_n)$ where

$$g_n = f \text{ or } g_n = Ef_n.$$

It is a much more difficult task to investigate the question of asymptotic normality of statistics of the form

$$(\varphi[d(f_n, g_n)] - a_n)/b_n$$

where $\{a_n\}$ and $\{b_n\}$ are suitable centralizing and normalizing deterministic sequences.

Typical examples of such statistics are built with p^{th} powers $(\varphi(z) = z^p)$ of L_p-errors. The L_p-error $(1 \leqslant p < \infty)$ between f_n and g_n is defined by

$$\|f_n - g_n\|_{p,\,\omega} = \left\{ \int_{\mathbb{R}} |f_n(x) - g_n(x)|^p \, \omega(x) \, \mathrm{d}\lambda(x) \right\}^{1/p},$$

where the weight function $\omega(.)$ is nonnegative, possibly random and/or depending on n.

Among L_p-errors $(1 \leqslant p \leqslant \infty)$ three are very popular. They are the L_1, L_2 and L_∞ errors. The last one is defined, with the same notation as above, by

$$\|f_n - g_n\|_{\infty,\,\omega} = \sup_{x \in \mathbb{R}} |f_n(x) - g_n(x)| \, \omega(x).$$

The L_1-error with $\omega(.) \equiv 1$ is certainly the more natural error in the context of density estimation because it is well defined whenever f_n and g_n are integrable and invariant under measurable one-to-one onto transformations (like rescaling of the axes) and can be expressed by means of measures of sets. The L_1- and L_∞-errors have the nice property of being easily visualized. The L_1-distance between two curves is the measure of the area between these curves whereas the L_∞-distance between them is the width of the smallest band of plane containing both of them. This visual aspect is important when using graphical tools in exploratory data analysis. The L_2-error is known for its nice properties allowing considerable specific developments. Section 4 will exploit the possibility of writing it as a scalar product. Many distances or pseudo-distances can be put in place of $d(f_n, g_n)$ depending on the problem under study and the aspects to be emphasized. We will illustrate the last section of the paper by considering the Kullback-Leibler divergence. Some recent works have shown the need for convergence results and CLT for pseudo-distances inducing stronger topologies than L_p-distances (for instance in the area of communication networks, see Györfi, Liese, Vajda and van der Meulen (1998) and Berlinet, Vajda and van der Meulen (1998)).

3. Strong approximation

As a rich literature is available on approximation of empirical processes (see Shorack and Wellner (1986), Wellner (1992) for a review of applications and van der Vaart and Wellner (1996)), a natural idea is to express the difference $(f_n(x) - f(x))$ as a function of the standard empirical process

$$\alpha_n(t) = \sqrt{n}(F_n(t) - F(t)), \qquad t \in \mathbb{R}.$$

Indeed, we can write

$$f_n(x) - f(x) = (f_n(x) - Ef_n(x)) + (Ef_n(x) - f(x))$$

and note that

$$f_n(x) = \int_{\mathbb{R}} K_n(x,t)\,\mathrm{d}F_n(t)$$

and

$$Ef_n(x) = \int_{\mathbb{R}} K_n(x,t)\,\mathrm{d}F(t).$$

The term $(Ef_n(x) - f(x))$ is non random. Suitable analytical conditions on the generalized kernel K_n, on the density f and the weight function ω will make it tend to zero. Now,

$$f_n(x) - Ef_n(x) = n^{-1/2} \int_{\mathbb{R}} K_n(x,t)\,\mathrm{d}\alpha_n(t).$$

We know that the asymptotic behavior of the empirical process $\alpha_n(.)$ is close to the behavior of $B_n(F(.))$ where $\{B_n(u), 0 \leqslant u \leqslant 1\}$ is a sequence of Brownian bridges: $\{B_n(u), 0 \leqslant u \leqslant 1\}$ has the same distribution as $\{W(u) - uW(1), 0 \leqslant u \leqslant 1\}$ where W is the standard Brownian motion on $(0,1)$. $B_n(.)$ is a Gaussian process with

$$E(B_n(t)) = 0 \qquad \text{and} \qquad E(B_n(s)B_n(t)) = \min(s,t) - st, \qquad 0 \leqslant s,t \leqslant 1.$$

This approximation result applies well to errors that are functionals of $(f_n(.) - g_n(.))$ like L_p-errors $(1 \leqslant p \leqslant \infty)$. Let

$$I_n(p) = \|f_n - Ef_n\|^p_{p,\,\omega}$$

and let $\hat{I}_n(p)$ be obtained from $I_n(p)$ by substituting

$$n^{-1/2} \int_{\mathbb{R}} K_n(x,t)\, dB_n(F(t)) \qquad \text{for} \qquad (f_n(x) - Ef_n(x)).$$

To get the desired result, that is the convergence in distribution to the standard normal $\mathscr{N}(0,1)$ of

$$(I_n(p) - a_n)/b_n,$$

we have to prove both the convergence

$$(\hat{I}_n(p) - a_n)/b_n \xrightarrow{\mathscr{D}} \mathscr{N}(0,1) \tag{5}$$

and the asymptotic negligibility (convergence in probability to zero) of

$$(I_n(p) - \hat{I}_n(p))/b_n. \tag{6}$$

The convergence of $(\hat{I}_n(p) - a_n)/b_n$ is proved by using moments and covariance properties of Brownian bridges together with a standard CLT. The asymptotic negligibility of $(I_n(p) - \hat{I}_n(p))/b_n$ is proved by applying strong approximation theorems on the empirical process.

However, this short presentation should not give the reader the feeling that things are easy. Heavy calculations are needed to handle stochastic integrals appearing in both parts of the proof (see for instance Horváth, 1991). The main drawback of the technique is the necessity, up to now, to assume analytical conditions on f or on the smoothing parameter leading to suboptimal rates of convergence.

Hall (1984) proposed a new method of proof to deal with nonparametric estimators of a multivariate density and to include the case of optimally constructed estimators, not covered in previous papers. While this method works for the general estimates defined by (3) it is limited to the integrated square error ($p = 2$)

$$I_n(2) = \|f_n - f\|_{2,\,\omega}^2 = \int_{\mathbb{R}} (f_n(x) - f(x))^2\, \omega(x)\, \mathrm{d}\lambda(x).$$

The specific treatment of $I_n(2)$ relies on the possibility of writing it as the sum of three terms:

$$\begin{aligned} I_n(2) = {} & \|f_n - Ef_n\|_{2,\,\omega}^2 \\ & + 2\int_{\mathbb{R}} (f_n(x) - Ef_n(x))(Ef_n(x) - f(x))\, \omega(x)\, \mathrm{d}\lambda(x) \\ & + \|Ef_n - f\|_{2,\,\omega}^2. \end{aligned}$$

The last term is deterministic and can be analyzed by analytic methods (this needs regularity assumptions on f).

The second term is readily described by a CLT because it can be written as a sum of i.i.d. random variables.

The first term may be expressed as

$$\begin{aligned} \|f_n - Ef_n\|_{2,\,\omega}^2 = 2 & \sum_{1\leqslant i<j\leqslant n} H_n(X_i, X_j) \\ & + \sum_{i=1}^{n} \int_{\mathbb{R}} (K_n(x, X_i) - EK_n(x, X_i))^2\, \omega(x)\, \mathrm{d}\lambda(x), \end{aligned}$$

where

$$H_n(x, y) = \int_{\mathbb{R}} (K_n(u, x) - EK_n(u, X_1))(K_n(u, y) - EK_n(u, X_1))\, \omega(u)\, \mathrm{d}u.$$

The second term in the expression of $\|f_n - Ef_n\|_{2,\omega}^2$ is a sum of i.i.d. random variables and therefore satisfies a CLT. The first term equals twice a centered U-statistic with a variable kernel function H_n. This means that H_n is a symmetric function satisfying $E(H_n(X_1, X_2)) = 0$. Moreover, this U-statistic is degenerate in the sense that

$$E(H_n(X_1, X_2) \mid X_1) = 0, \text{ a.s.}$$

Defining

$$Y_i = \sum_{j=1}^{i-1} H_n(X_i, X_j), \quad S_i = \sum_{j=2}^{i} Y_j, \ 2 \leqslant i \leqslant n,$$

and noting that

$$E(Y_i \bigm| X_1, \ldots, X_{i-1}) = 0, \ \text{a.s.}\ 2 \leqslant i \leqslant n,$$

one concludes that the sequence $\{S_i, 2 \leqslant i \leqslant n\}$ is a martingale in which

$$S_n = \sum_{1 \leqslant i < j \leqslant n} H_n(X_i, X_j).$$

Thus, using Brown's martingale CLT, Hall proves in his paper the following theorem for centered, degenerate U-statistics with a variable kernel.

Theorem 1. *Assume that H_n is symmetric, $E(H_n(X_1, X_2)|X_1) = 0$ a.s., $E(H_n^2(X_1, X_2)) < \infty$ for each n and define*

$$G_n(x, y) = E(H_n(X_1, x)H_n(X_1, y)).$$

If, as $n \longrightarrow \infty$,

$$(E(G_n^2(X_1, X_2)) + n^{-1}E(H_n^4(X_1, X_2)))/(E(H_n^2(X_1, X_2)))^2 \longrightarrow 0,$$

then

$$n^{-1} \sum_{1 \leqslant i < j \leqslant n} H_n(X_i, X_j)$$

is asymptotically normally distributed with zero mean and variance $\frac{1}{2}E(H_n^2(X_1, X_2))$.

Finally, $I_n(2)$ can be written as the sum of a deterministic term $\|Ef_n - f\|_{2,\omega}^2$ and of three random terms. Applying twice a standard CLT together with Theorem 1 and a suitable approximation result for $\|Ef_n - f\|_{2,\omega}^2$ one gets asymptotic normality for $I_n(2)$ in the case where one of the random term is dominant. Otherwise one has to prove joint asymptotic normality to be able to conclude for $I_n(2)$.

The remark made at the end of Section 3 concerning undesirable assumptions on f remains valid.

5. Poissonization

The word "Poissonization" covers essentially two techniques. The first consists in randomizing a sample size by means of a Poisson random variable while the second exploits the possible representation of a multinomial vector by a vector of independent Poisson random variables conditionally on their sum. Their interest relies on the nice properties of Poisson processes, namely independence of their increments and the behavior of their moments. These properties considerably simplify calculations. In many situations, Poisson approximations of empirical processes have better rates of convergence than Gaussian ones. Their superiority is a consequence of their better behavior on the tails and can be amplified in the case of weighted empirical processes.

5.1. Poissonization of sample size.

The idea goes back to Kac (1949), of approximating

$$nF_n(t) = \sum_{i=1}^{n} 1_{\{X_i \leqslant t\}}$$

by

$$A_n(t) = \sum_{i=1}^{N_n} 1_{\{X_i \leqslant t\}},$$

where N_n is a Poisson random variable with mean n, independent of $\mathbf{X} = (X_1, \ldots, X_n, \ldots)$. The representation of the process nF_n by means of A_n plus a remainder R_n is known as the Kac representation. The process A_n is a Poisson process with intensity $E(A_n(t)) = nF(t)$ and, if the X_i's are supposed to be uniform on $(0, 1)$, we have as $n \longrightarrow \infty$ the following approximation result in the sense of weak convergence (see Csörgő and Horváth, 1993):

$$n^{-1/4}\{nF_n(t) - A_n(t) + t(N_n - n)\} \longrightarrow (\text{sign}Z)|Z|^{1/2}B(t)$$

where $B(.)$ is a Brownian bridge and Z is a standard $\mathcal{N}(0, 1)$ random variable independent of $B(.)$. The sample size Poissonization can be used not only for empirical processes but also for any statistics built from $(X_1, X_2, \ldots, X_n)$. As an example let us mention the pioneering work of Rosenblatt (1975) exploiting this method initially suggested by Bickel to study the behavior of bivariate kernel density estimates

$$f_n(x) = \frac{1}{nh_n^2} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right),$$

with a kernel K supposed to be zero outside a compact set of $\mathbb{R}^2$. In his paper, Rosenblatt considered

$$f_n^*(x) = \frac{1}{nh_n^2} \sum_{i=1}^{N_n} K\left(\frac{x - X_i}{h_n}\right),$$

where N_n is a Poisson(n) random variable independent of $\mathbf{X}$, as a tool to prove the asymptotic normality of the square of the L_2-error $\|f_n - f\|_{2,\omega}^2$. Using the properties of the Poisson process A_n, one can write

$$\|f_n^* - Ef_n^*\|_{2,\omega}^2 = \sum_{j,k} U_{j,k}(n)$$

where $U_{j,k}(n)$ and $U_{j',k'}(n)$ are independent if $|j - j'| \geqslant 2$ or $|k - k'| \geqslant 2$. Then a CLT can be proved for $\|f_n^* - Ef_n^*\|_{2,\omega}^2$ just as for an m-dependent process using the Liapunov condition. The asymptotic normality of the L_2-error associated with f_n^* is then transferred to f_n by means of approximation theorems. Rosenblatt proved his results in the multivariate case in which this methodology works as well, underlining the fact that the Poissonization technique requires weaker assumptions on the underlying density and the parameters of estimates than the strong Gaussian approximation (see Bickel and Rosenblatt, 1973). The same technique was used by Horváth (1991) to prove CLTs for L_p norms of errors associated with the above density estimates. It should be stressed that, as N_n is unbounded, the effective computation of an estimate based on $(X_1, X_2, \ldots, X_{N_n})$ would require the knowledge of the values of the infinite sequence of variables $\mathbf{X}$. Therefore Poissonized samples and more generally samples with unbounded random size are mainly probabilistic tools.

5.2. Poisson representation of multinomial distribution.

The distribution of a multinomial vector $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_k)$ with parameters $n, p_1, \ldots, p_k$ can be represented as the distribution of a vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_k)$ of k independent Poisson variables with expected values $np_1, \ldots, np_k$ and subject to the condition $Y_1 + Y_2 + \ldots + Y_k = n$. The Poissonization technique exploits this property and is widely used in urn occupancy problems to obtain asymptotic formulas for sequential occupancy probabilities or expected values of waiting times. See Johnson and Kotz (1977) or Barbour, Holst and Janson (1992) and the references therein. As a number of test statistics (most notably chi-square and likelihood ratio statistics) have the form

$$S_k = \sum_{i=1}^{k} \psi_i(Z_i),$$

where $\psi_1, \dots, \psi_k$ are measurable real-valued functions, it is fundamental to give conditions under which the limiting distribution of S_k is known. The first results of this kind were given by Steck (1957) using the Poisson representation of $\mathbf{Z}$, and extended by Morris (1975). A new area of application for this type of Poissonization was recently opened in functional estimation to get limit theorems for functions of global errors $\varphi[d(f_n, g_n)]$ (see Berlinet (1995) for formalization of the method and list of references). Different sorts of nonparametric estimates are defined as functions of multinomial vectors $\mathbf{Z} = (Z_1, \dots, Z_k)$ defined from partitions of $\mathbb{R}$, each Z_i being a number of observations falling into some cell of a partition (think of the standard histogram or regressogram for example). The idea to get the asymptotic distribution of $\varphi[d(f_n, g_n)]$ is to replace vectors $\mathbf{Z}$ by vectors $\mathbf{Y}$ of Poisson variables conditioned on their sum and then to transfer back the results to the original statistics, approximated by functions of multinomial variables. The key theorem for this purpose was proved by Beirlant, Györfi and Lugosi (1994). It relies on the idea of partial inversion for obtaining characteristic functions of conditional distributions. Roughly speaking their theorem states conditions guaranteeing that whenever a CLT holds for statistics of the form

$$\sum_j \psi_j(Y_j)$$

then it also holds for the statistics

$$\sum_j \psi_j(Z_j),$$

where functions and variables depend on n. For details and application to various kinds of estimates and different types of errors, see Beirlant, Györfi and Lugosi (1994), Beirlant and Mason (1994), Berlinet, Devroye and Györfi (1995) and Berlinet (1995). Now, suppose that we know a density g such that the Kullback-Leibler divergence between f and g, i.e.

$$D(f, g) = \int_{\mathbb{R}} f(x) \ln \frac{f(x)}{g(x)} \, \mathrm{d}\lambda(x)$$

is finite. Consider the Barron estimate f_n of f defined in Section 2. Beirlant, Györfi and van der Meulen (1992) proved that under the conditions

$$\lim_{n\to\infty} h_n = 0, \ \lim_{n\to\infty} nh_n = \infty, \tag{7}$$

we have

$$\lim_{n\to\infty} D(f, f_n) = 0 \ \text{ a.s.}$$

and

$$\lim_{n\to\infty} E(D(f, f_n)) = 0.$$

The question of asymptotic normality of

$$D(f, f_n) - ED(f, f_n)$$

has been very recently investigated by Berlinet, Györfi and van der Meulen (1997) using Poissonization techniques. Let $\mu^*_{N_n}$ be the Poissonized empirical measure defined on Borel sets by

$$\mu^*_{N_n}(A) = \frac{1}{n}\sum_{i=1}^{N_n} I_{X_i \in A}.$$

If the functions ψ_j are chosen as

$$\psi_j(x) = n\sqrt{2h_n}\frac{n^2\mu(A_{nj})}{2(n\mu(A_{nj})+1)^2}\left((x-\mu(A_{nj}))^2 - E(\mu^*_{N_n}(A_{nj}) - \mu(A_{nj}))^2\right)$$

$(j = 1, \dots, m_n)$, then

$$n\sqrt{2h_n}(D(f, f_n) - ED(f, f_n)) = \sum_{j=1}^{m_n} \psi_j(\mu_n(A_{nj})) + \Delta_n,$$

where Δ_n is shown to tend to zero in probability. Hence the problem is reduced to the proof of a CLT for

$$S_n = \left(t\sum_{j=1}^{m_n} \psi_j(\mu_{N_n}(A_{nj}))\right) + v\frac{N_n - n}{\sqrt{n}}.$$

This is done by using moment properties of Poisson variables and the Liapunov condition. In this way Berlinet, Györfi and van der Meulen (1997) obtained the following result, which is the first asymptotic distributional theorem for the Kullback-Leibler divergence of density estimates.

Theorem 2. *Let μ and ν be probability measures on $\mathbb{R}^d$ with densities f and g with respect to the Lebesgue measure. Let $\overline{S}_\mu$ be the closure of the set $S_\mu = \{x\colon\ f(x) \neq 0\}$ and let f_n be given by*

$$f_n(x) = (n\mu_n(A_n(x)) + 1)a_n g(x)$$

with (h_n) satisfying

$$\lim_{n\to\infty} h_n = 0, \ \lim_{n\to\infty} nh_n = \infty.$$

If $D(f, g) < \infty$ and $\nu(\overline{S}_\mu - S_\mu) = 0$ then

$$n\sqrt{2h_n}[D(f, f_n) - E(D(f, f_n))] \xrightarrow{\mathscr{D}} \mathscr{N}(0, \sigma^2)$$

as $n \to \infty$, where $\sigma^2 = \nu(S_\mu) > 0$.

Observe that the asymptotic variance is less than or equal to 1, in any dimension, for all densities f for which the consistency in Kullback-Leibler divergence of the estimate is guaranteed.

The problem of strong consistency in the sense of information divergences

$$D_\varphi(f,g) = \int g\,\varphi\left(\frac{f}{g}\right)\,\mathrm{d}\lambda$$

topologically stronger than the Kullback-Leibler divergence is more difficult. See in this respect Györfi, Liese, Vajda and van der Meulen (1998) who considered the χ^2-divergence ($\varphi(t) = |1-t|^2$), and Berlinet, Vajda and van der Meulen (1998) for other divergences. It is an open problem to find the asymptotic distribution of errors defined from general φ-divergences.

The representation of multinomial vectors by vectors of i.i.d. Poisson variables given their sum can be viewed as a particular case of the sample size Poissonization. Let $\mathbf{X} = \{X_i\colon\ i = 1, 2, \ldots\}$ be a sequence of i.i.d. random variables and let $\mathscr{P} = \{A_1, A_2, \ldots, A_k\}$ be a partition of $\mathbb{R}$. Consider a Poisson random variable N_n, with mean n and the Poissonized empirical measure $\mu^*_{N_n}$. The vector $(n\mu^*_n(A_1), \ldots, n\mu^*_n(A_k))$ is a vector of independent Poisson random variables with the same distribution as $(n\mu_n(A_1), \ldots, n\mu_n(A_k))$, conditionally on their sum.

To obtain limit theorems, one often has to combine different methods. To illustrate this, let us mention the result given by Berlinet, Devroye and Györfi (1995) for the standard histogram f_n built from a partition $\mathscr{P}_n$ of $\mathbb{R}$ into intervals A_{nj}, $j \geqslant 1$, with equal measure $h_n = cn^{-1/3}$. For any continuously differentiable density f,

$$\sqrt{n}\left(\|f_n - f\| - E\|f_n - f\|\right)/\sigma$$

is asymptotically Gaussian $\mathscr{N}(0,1)$. The constant σ is given explicitly and shown to be less than $1 - 2/\pi$. The authors prove the result with a centering constant equal to $E\|f^*_n - f\|$, where f^*_n is the Poissonized histogram defined by

$$f^*_n(x) = \frac{\mu^*_{N_n}(A_n(x))}{h_n}.$$

For the statistic $(\|f_n - f\| - E\|f^*_n - f\|)$ the technique of replacement exposed in this paragraph is used. It remains to prove that the Poissonization of sample size provides a suitable approximation f^*_n to f_n. This is done through careful inspection of the L_1-error.

6. By way of conclusion

As we have mentioned, there are prima facie reasons why, up to now, strong approximation methods need additional assumptions: the computation and the approximation of the stochastic integrals involved in these methods use regularity or boundedness properties of the unknown density. These assumptions cannot be checked from a data set and therefore are not desirable in view of statistical applications. The same is true for methods needing a good approximation of f by Ef_n. As suggested early by Rosenblatt for quadratic functionals, it seems that the technique of Poissonization is more natural, in the sense that it requires weaker assumptions on the underlying density and the parameters of the estimates. Other sample size randomization techniques appeared in the literature (Pollard, 1982). To obtain limit theorems under weak assumptions it seems that some standard analytical tools should be replaced by more probabilistic methods relying on probabilities of sets.

References

[1] *Barbour, A. D., Holst, L. and Janson, S.*: Poisson Approximation. University Press, Oxford, 1992.

[2] *Barron, A. R.*: The convergence in information of probability density estimators. Presented at IEEE ISIT, Kobe, Japan, June 19–24, 1988.

[3] *Barron, A. R., Györfi, L. and van der Meulen, E. C.*: Distribution estimates consistent in total variation and in two types of information divergence. IEEE Trans. on Information Theory *38* (1992), 1437–1454.

[4] *Beirlant, J., Györfi, L. and Lugosi, G.*: On the asymptotic normality of the L_1 and L_2 errors in histogram density estimation. Canadian Journal of Statistics *3* (1994), 309–318.

[5] *Beirlant, J. and Mason, D. M.*: On the asymptotic normality of L_p norms of empirical functionals. Mathematical Methods of Statistics *4* (1994), 1–15.

[6] *Berlinet, A.*: Central limit theorems in functional estimation. Bulletin of the International Statistical Institute *56* (1995), 531–548.

[7] *Berlinet, A., Devroye, L. and Györfi, L.*: Asymptotic normality of L_1 error in density estimation. Statistics *26* (1995), 329–343.

[8] *Berlinet, A., Györfi, L. and van der Meulen, E. C.*: Asymptotic normality of relative entropy in multivariate density estimation. Revue de l'Institut de Statistique de l'Université de Paris *41* (1997), 3–27.

[9] *Berlinet, A., Vajda, I. and van der Meulen, E. C.*: About the Asymptotic Accuracy of Barron density estimates. Trans. IEEE on Inform. Theory *44* (1998), 999–1009.

[10] *Bickel, P. and Rosenblatt, M.*: On some global measures of the deviation of density function estimates. The Annals of Statistics *1* (1973), 1071–1095.

[11] *Csörgő, M. and Horváth, L.*: Weighted Approximations in Probability and Statistics. Wiley, New York, 1993.

[12] *Györfi, L., Liese, F., Vajda, I. and van der Meulen, E. C.*: Distribution estimates consistent in χ^2-divergence. Statistics *32* (1998), 31–58.

[13] *Hall, P.*: Central limit theorem for integrated square error of multivariate nonparametric density estimators. Journal of Multivariate Analysis *14* (1984), 1–16.

[14] *Heyde, C. C.*: Central limit theorem. Encyclopedia of Statistical Sciences *4* (1983), 651–655.

[15] *Horváth, L.*: On L_p norms of multivariate density estimators. The Annals of Statistics *19* (1991), 1933–1949.

[16] *Johnson, N. L. and Kotz, S.*: Urn Models and their Application. Wiley, New York, 1977.

[17] *Kac, S.*: On deviations between theoretical and empirical distributions. Proceedings of The National Academy of Sciences of USA *35* (1949), 252–257.

[18] *Morris, C.*: Central limit theorems for multinomial sums. The Annals of Statistics *3* (1975), 165–188.

[19] *Pollard, D.*: Beyond the Heuristic Approach to Kolmogorov-Smirnov Theorems. Essays in Statistical Science. Festschrift for P. A. P. Moran (J. Gani and E. J. Hannan, eds.). Applied Probability Trust, 1982.

[20] *Rosenblatt, M.*: A quadratic measure of deviation of two-dimensional density estimates and a test of independence. The Annals of Statistics *3* (1975), 1–14.

[21] *van der Vaart, A. W. and Wellner, J. A.*: Weak Convergence and Empirical Processes with Applications to Statistics. Springer, New York, 1996.

[22] *Shorack, G. R. and Wellner, J. A.*: Empirical Processes with Applications to Statistics. Wiley, New York, 1986.

[23] *Steck, G. P.*: Limit theorems for conditional distributions. University of California Publications in Statistics.

[24] *Wellner, J. A.*: Empirical processes in action: a review. International Statistical Review *60* (1992), 247–269.

Author's address: *Alain Berlinet*, Department of Mathematics, University of Montpellier, France, e-mail: `berlinet@helios.ensam.inra.fr`.