Comment

Ron Pyke

I have greatly enjoyed reading this paper by David Pollard. It is a further example of his fine expositing skills. By focusing on two particular problems, he underlines for statisticians the practical values that are intrinsic to the subject of weak convergence of empirical processes.

It is slightly more than 60 years since Harald Cramér introduced the idea of an empirical distribution function for real random variables and suggested its use in statistics (Cramér, 1928). Shortly thereafter, the Glivenko-Cantelli-Kolmogorov result of 1931 showed that the empirical distribution function was a strongly consistent estimator of the population distribution function. This was followed for about 25 years by a virtual explosion of applied and theoretical activity on distribution-free nonparametric procedures: Kolmogorov-Smirnov and Cramér-von Mises type statistics: one-sided and two-sided; one-sample and two-sample; weighted and unweighted; asymptotic and exact results; with tables of critical values provided for most.

Forty years ago, in the midst of this activity, J. L. Doob proposed a result that would enable one to obtain the asymptotic behavior of most of the procedures that had been, or would ever be, proposed from an appropriate limiting Gaussian process, a tied-down Brownian motion. (Cf. Doob, 1949). Out of this heuristic beginning, a vast literature has emerged concerning the asymptotic behavior of empirical processes. Throughout this research, the central and enabling property of the empirical measure of a sample of iid observations $X_1, X_2, \cdots, X_n$ has been its basic structure as a *sample average* of iid objects, namely,

$$P_n = n^{-1}(\delta_{X_1} + \delta_{X_2} + \cdots + \delta_{X_n})$$

where δ_x is the degenerate probability measure that puts probability 1 at x. Because of this structure, it is natural that the asymptotic distributional, or weak convergence, results are referred to as central limit theorems (CLT) for empirical processes. (It also suggests interest in other sample average limit laws for P_n, such as the SLLN and LIL.) Shortly before (1), Pollard states that, "In some asymptotic sense, the process $\nu_n f(\cdot, t)$ is approximately Gaussian." This is as close as the author gets to mentioning a CLT for empirical processes. This brief sentence encompasses an enormous literature that pertains to the asymptotic distribution of empirical processes.

Although the setting for these CLT's is much more general than for the classical CLT and the technical aspects are accordingly more complex, their much greater applicability to statistics makes their study worthwhile. In fact, I view any CLT for empirical processes as a conveniently packaged collection of many individual limit theorems of importance to statisticians. For exposition purposes to statisticians, I prefer to define convergence in law of empirical processes (i.e., when a CLT holds) to mean precisely the convergence in law of all statistics that are continuous functions of the empirical process. (Cf. Pyke and Shorack, 1968). From this viewpoint, CLT's for empirical processes are powerful tools that statisticians, or their consulting probabilists, can check out of our Asymptotic Methods' Toolroom. Often, however, these tools need to be individually customized to handle statistics that are only approximately continuous, and this is the situation with the two examples presented here by David Pollard; the simple substitution of $\bar{X}$ for t in the first problem is unfortunately one complication that requires technical care to justify the natural Taylor-expansion heuristics, while the question of asking for the location of a min or max as in the second problem is in and of itself another complication. Regardless of the excellent quality of exposition, the level of this complexity cannot be hidden. The important message, however, is that applications of this type *can* be handled by the theory, regardless of whether or not the particular methodology is understood or even fully appreciated by the user.

Major advances in the theory of statistics are driven ultimately by applications. In 1978, I felt that rather complete results about all three major types of limit theorems for empirical processes were available; namely, for the CLT or weak convergence, Dudley (1978); for the SLLN or Glivenko-Cantelli result, Steele (1978); and for the LIL, Kuelbs and Dudley (1980; a preprint was available in 1978). I therefore used my 1978 IMS Special Invited Lecture to survey the 50 years since Cramér (1928) and to encourage that the rather complete theory then available be brought to bear on applications of empirical process involving multidimensional data. The considerable theoretical advances of the last decade clearly indicate that the subject's theory and methodologies were far from complete in 1978. Many major advances have occurred since then, and along with these have come

Ron Pyke is Professor, Department of Mathematics, University of Washington, Seattle, Washington 98195.

several fine monographs and survey papers, including the one under discussion. These expositions make the theory and its application much more accessible today for study and implementation, even though much more can still be done to improve the formulation of results so as to facilitate their applicability. Ironically, it is the applications that point the way to such improvements.

Since 1978, relatively few researchers have ventured in the direction of proposing and evaluating empirical process applications, particularly ones concerning inferences involving multidimensional data. David Pollard has been one of the most successful through his contributions to applications involving kernel density estimators, U-statistics, econometrics, regression, k-means clustering, and location estimators.

In the present paper, both problems used by Pollard to illustrate aspects of these asymptotic methods involve statistics that are expressible in terms of the translates of a fixed function h; i.e., $f(x, t) = h(x - t)$ where in the first problem, $h(y) = |y|$, and in the second problem, $h(y) = \min\{1, |y|^2\}$. (As an aside, I wonder if $h(y) = |y|^2/(1 + |y|^2)$ is a reasonable smooth substitute, or even $h(y) = 1 - \exp(-|y|^2)$; this is of course not an implied criticism, since criticism was forestalled by the legitimate *caveat* stated in the introduction of the second problem!) Although the family of all functions that can be obtained as translates of a fixed function might seem to be rather restrictive and uninteresting, it might be well to emphasize that in many natural cases it is actually full enough to determine the underlying probability measure itself. Let me elaborate.

Procedures based on empirical processes can be thought of as goodness-of-fit procedures; $P_n - P$ is indeed the difference between the observed and the expected. In particular, if A is a subset of the sample space, $P_n(A) - P(A)$ is in fact the difference between the observed proportion of observations in A and its expected value under P. Similarly, for an integrable function f, $P_n(f) - P(f)$ is the difference between the observed sample average $n^{-1}(f(x_1) + \cdots + f(X_n))$ and its expectation $Ef(x)$. If the sample space were to be partitioned into a finite number of sets A, the classical Chi-square statistic would give one measure of the distance between P_n and P. If the data is Euclidean, other distance measures can be considered, including the usual Kolmogorov-type distance

$$D_n = \sup_x |P_n((-\infty, x]) - P((-\infty, x])|$$

and the Cramér-von Mises distance

$$W_n^2 = \int |P_n((-\infty, x]) - P((-\infty, x])|^2 \, dP(x).$$

Each of these involves the family of lower orthants and this family is simply the family of sets one gets by translating a particular orthant, say $(-\infty, 0]$. If we write $\| P_n - P \|_{\mathscr{A}}$ for the supremum over $A \in \mathscr{A}$ of $|P_n(A) - P(A)|$, then D_n is just this sup-metric when $\mathscr{A}$ is the family of lower orthants. However, in multidimensional situations, orthants are far from natural, and other more attractive possibilities exist. The Cramér-Wold result states that $\| P_n - P \|_{\mathscr{A}}$ is a distance when $\mathscr{A}$ is the family of half-spaces; this family is formed from a single half-space by making all translations *and* rotations. Recent work by Beran and Millar (1986) shows the applicability and value of this family for obtaining confidence sets for multidimensional distributions.

Less well known is the fact due to Sapogov (1974) (cf. Pyke, 1984) that the collection of all translates of a fixed bounded set of positive Lebesgue measure is also a determining class; that is, if two probability measures agree on the class, they must be equal. For example, if two probability distributions agree on every ball of radius 1, then the two distributions are equal. In such cases, $\| P_n - P \|_{\mathscr{A}}$ would again be an appropriate Kolmogorov-type statistic for measuring the goodness-of-fit of P to the data represented by P_n. A fairly extensive simulation study of tests based on these "scanning" statistics has been reported in Pyke and Wilbour (1988).

By identifying sets with their indicator functions, a family of sets formed by the translates of a fixed set becomes a special case of the class of translates of a given function. Suppose h is a given real valued function defined on $\mathbb{R}^d$ and let $\mathscr{T}_h = \{h(\cdot - x): x \text{ in } \mathbb{R}^d\}$ be the family of all of the translates of h. For many useful functions h, these translation or scanning families are determining classes in the sense that knowledge of all of the expectations determines the underlying distribution. Here is an example: *If h is non-negative and its Fourier transform, namely,*

$$\hat{h}(u) = \int \exp(iu \cdot x) h(x) \, dx,$$

is nonzero for a dense set of u in $\mathbb{R}^d$, then whenever X and Z are any two r.v.'s in $\mathbb{R}^d$ for which

$$Eh(X - x) = Eh(Z - x) \quad \textit{for all } x \textit{ in } \mathbb{R}^d,$$

we have that X and Z are identically distributed.

To prove this, simply multiply both sides of the identity by $\exp(iu \cdot x)$ and integrate over x. This gives the Fourier transform of a convolution and results in the identity

$$\hat{h}(-u) E(e^{iu \cdot Y}) = \hat{h}(-u) E(e^{iu \cdot Z}), \quad \text{for all } u \text{ in } \mathbb{R}^d.$$

Thus when the assumption about $\hat{h}$ holds, we can factor it out and conclude, as desired, that Y and Z have the same characteristic function, and hence the same distribution.

Translation families of functions arise, for example, in kernel function density estimation, in which the kernel is often a density function itself (e.g., the normal kernel) that satisfies the assumption in the above proposition. One should also note that the translations of $h(x) = \min(1, |x|^2)$, used in Pollard's second problem, form a determining class. This can be seen by applying the above proposition to $1 - h$; the constant function is invariant under translation. In fact, Pollard's problem seems to be easier to describe in terms of $g(x) = 1 - h(x) = \max(0, 1 - |x|^2)$ as a sort of "shell game." The function g is a parabolic shell and the game is to move it over the data in the plane until one finds the place where the score (the sum of the heights of the shell above the data) from the points covered by the shell, is a maximum. Pollard assumes that P is such that the expected score is itself approximately a parabola. By weak convergence, the difference between the observed and expected scores is approximately $n^{-1/2}$ times a Gaussian process. So as the observed score hugs the parabola-shaped expected score, the maximum observed score will appear near the place where the maximum expected score occurs, namely 0, and the difference between the two locations will depend on the behavior of the Gaussian process in the neighborhood of 0.

The example is exceptionally good. There are several approaches that can be used to obtain limiting distributions of statistics which are defined explicitly in terms of empirical processes. Different approaches are best in different situations; the more techniques one knows the better. One useful approach is to replace weak convergence by strong convergence but this approach also has its problems in this case. As more general weak convergence results (CLT's) for empirical processes have evolved, results showing that "weak implies strong" have kept pace. These results enable one to replace weak convergence by strong (pointwise) convergence, a substitution that has the effect of replacing many problems of a probabilistic nature with more standard problems in analysis. However, in this example, even though the nonzero part of g has a Taylor's expansion, $g(x - t) = g(x) - t\nabla g(x) +$ remainder, in which most of the time the gradient, $\nabla g(x)$, and the remainder are elementary functions, namely, $-2x$ and $|t|^2$, respectively, the problem is far from straightforward. If the mark of a good example is the degree to which it probes the applicability of theory and leads to reformulations that either expand or facilitate this applicability, then high marks are indeed in order here.

At the end of the paragraph preceding (3), the author "explains in part why traditional methods have difficulty with this (second) problem." I sense an implicit challenge here. Write $g(x) = (1 - |x|^2)^+$ and

$$d_{n,b}(x) = n^{1/2}\{g(x - n^{-1/2}b) - g(x)\}.$$

The minimizing of $H_n(t)$ over t is equivalent to the maximizing over b of

$$\begin{aligned} L_n(b) &= n\{H_n(0) - H_n(bn^{-1/2})\} \\ &= n\{H_n(0) - H(0) + H(0) - H(bn^{-1/2}) \\ &\qquad + H(bn^{-1/2}) - H_n(bn^{-1/2})\} \\ &= (\nu_n + n^{1/2}P)\, d_{n,b}. \end{aligned}$$

The author postulates that the second part is a quadratic, namely, $n^{1/2}Pd_{n,b} = -\tfrac{1}{2}|b|^2(1 + o(1))$. If we also assume with the author that it suffices to consider the maximization over $|b| < M$ for each M, we can partition the sample space into

$$B_1 = [|x| < 1 - n^{-1/2}M],$$

$$B_2 = [|x| > 1 + n^{-1/2}M]$$

and

$$B_3 = [|1 - |x|| \le n^{-1/2}M].$$

When $|b| < M$, $d_{n,b} = 0$ on B_2, $= 2x \cdot b + n^{-1/2}|b|^2$ on B_1 and is bounded by $4M$ on the annulus B_3 whose probability converges to $P[|X| = 1] = 0$. Thus

$$L_n(b) = 2Y_n \cdot b - |b|^2/2 + o(1)$$

where

$$Y_n = \nu_n(x1_{B_1}) \xrightarrow{L} Y, \quad \text{a } N(0, EX^2 1_{[|X|<1]}) \text{ r.v.,}$$

and where the remainder is uniform for $|b| < M$. Thus $\hat{b}_n$, the value for which $L_n(b)$ is maximized, and $2Y_n$, the value for which the quadratic $Q(b) = 2Y_n \cdot b - |b|^2/2$ is maximized, converge to the same limit, as desired. I leave to the reader the challenge of identifying where further details are needed and which traditional methods could be invoked, making use of direct analysis of $d_{n,b}$.

Let me be quick to emphasize that the value of the second example did not lie in its inability to be solved by traditional methods, but rather in the ease with which it can be handled by the new methods discussed in the paper; c.f. the short conclusion in (5.4). The value is really much greater since the particular problem is intended only to provide a simple illustration of the newer methods' broad applicability.

When h is such that the translation family $\mathscr{T}_h$ is a determining class of functions, a metric on the space of probability measures can be defined in terms of the sup-norm over $\mathscr{T}_h$, namely, $\|P - Q\|_{\mathscr{T}_h} = \sup\{|(P - Q)f| : f \in \mathscr{T}_h\}$. This in turn enables one to consider the Kolmogorov-type statistic

$$D_n(h, P) := \|P_n - P\|_{\mathscr{T}_h} = \sup_t |(P_n - P)h(\cdot - t)|$$

as a measure of the distance between P_n and P. As an example, let me phrase the author's second problem more explicitly as an estimation problem for a translation parameter θ, and make use of the Kolmogorov-type statistic $D_n(g, P)$ based on the shell function $g(x) = (1 - |x|^2)^+$. Let $\mathscr{P} = \{P^\theta: \theta \in \mathbb{R}^2\}$ be the translation family of probability measures defined by $P^\theta(A) = P(A - \theta)$. Suppose one wishes to estimate θ by the minimum distance estimator, $\hat{\theta}_n$, defined as that value of $\theta \in \mathbb{R}^d$ which minimizes the distance

$$D_n(g, P^\theta) = \sup_t | P_n g(\cdot - t) - Pg(\cdot - t - \theta) |.$$

Under the assumption that the true parameter is $\theta_0 = 0$, it appears that the asymptotic distribution of $\hat{\theta}_n$ may be the same as that for Pollard's estimate, $\hat{\tau}_n$, the value of t at which $P_n g(\cdot - t)$ is maximized, even though the minimization problems are different. Let me offer as a third test of the author's methodology the question of determining the limiting distribution of $n^{1/2}\hat{\theta}_n$. This type of problem is similar to one considered by Blackman (1955), except that he used a Cramér-von Mises distance rather than a Kolmogorov one; in Pyke (1970) this simpler problem was used to illustrate the applicability of the "weak implies strong" methodology mentioned above.

Although I have directed my comments on the paper towards statisticians as users of this theory, I would stress that the paper is also of great value to those doing research in the area. From both viewpoints I greatly appreciate the efforts of David Pollard for preparing this valuable exposition.

ADDITIONAL REFERENCES

BERAN, R. and MILLER, P. W. (1986). Confidence sets for a multivariate distribution. *Ann. Statist.* **14** 431–443.

BLACKMAN, J. (1955). On the approximation of a distribution function by an empiric distribution. *Ann. Math. Statist.* **26** 256–267.

CRAMÉR H. (1928). On the composition of elementary errors. II. *Skand. Aktuarietidskr.* **11** 141–180.

DOOB, J. L. (1949). Heuristic approach to the Kolmogorov-Smirnov theorems. *Ann. Math. Statist.* **20** 393–403.

KUELBS, J. and DUDLEY, R. M. (1980). Loglog laws for empirical measures. *Ann. Probab.* **8** 405–418.

PYKE, R. (1970). Asymptotic results for rank statistics. In *Proc. First Symp. on Non-Parametric Techniques* (M. Puri, ed.) 21–40. Cambridge Univ. Press, Cambridge.

PYKE, R. and SHORACK, G. (1968). Weak convergence of a two sample empirical process and a new approach to Chernoff-Savage theorems. *Ann. Math. Statist.* **39** 755–771.

PYKE, R. and WILBOUR, D. C. (1988). New approaches for goodness-of-fit tests for multidimensional data. In *Statistical Theory and Data Analysis II* (K. Matusita, ed.) 139–154. North-Holland, Amsterdam.

SAPOGOV, N. A. (1974). A uniqueness problem for finite measures in Euclidean spaces. Problems in the theory of probability distributions. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov.* (*LOMI*) **41** 3–13. (In Russian.)

STEELE, J. M. (1978). Empirical discrepancies and subadditive processes. *Ann. Probab.* **6** 118–127.

Comment

Miklós Csörgő and Lajos Horváth

It is a pleasure to congratulate David Pollard on his masterly glimpse into the theory of empirical processes. His artful development here of the technique of Gaussian symmetrization, of the resulting maximal inequalities for Gaussian processes and their application in the empirical process context leaves us no room for comment on his methods, which extend the concept of a Vapnik-Červonenkis class of sets. He demonstrates the efficiency of these methods by use of two motivating, nontrivial asymptotic problems and succeeds very well in conveying the look and feel of a powerful tool of contemporary mathematical statistics.

There are also other powerful contemporary tools available for tackling asymptotic problems of mathematical statistics. The ones we have in mind are strong and weak approximations (almost sure and in probability invariance principles) for empirical and partial sum processes based on various forms of the Skorohod embedding scheme, or on various forms of the Hungarian construction. The quoted book of Shorack and Wellner (1986) is also an excellent source of information on these methods. For further references on the methods and their applications, we mention the books of Csörgő and Révész (1981), Csörgő (1983), and Csörgő, Csörgő and Horváth [CsCsH] (1986). For an insightful overview of strong and weak approximations we refer to Philipp (1986) (cf. also the review of Csörgő (1984)). Concerning Hungarian constructions, for those who are really interested, the papers of Bretagnolle and Massart (1989), and Einmahl (1989) are most recommended readings.

Here *we* make use of the first problem discussed by David Pollard to *illustrate* what we mean by *strong*

Miklós Csörgő is Professor of Mathematics and Statistics at Carleton University. His address is Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada K1S 5B6. Lajos Horváth is Associate Professor, Department of Mathematics, University of Utah, Salt Lake City, Utah 84112.