CHAPTER 1

ROBUST ESTIMATION VIA GENERALIZED L-STATISTICS: THEORY, APPLICATIONS, AND PERSPECTIVES

ROBERT SERFLING

University of Texas at Dallas

Abstract: Generalized L-statistics, introduced in Serfling (1984) and including classical U-statistics and L-statistics, are linear functions based on the ordered evaluations of a kernel over subsets of the sample observations. In particular, generalized median statistics fall within this class and are found to fulfill an interesting and potent principle, that "smoothing" followed by "medianing" yields a very favorable combination of efficiency and robustness. Extensive asymptotic theory now available for generalized L-statistics is reviewed, including asymptotic normality, strong convergence, large deviation, sequential fixed-width confidence interval, jackknife, and bootstrap results, as well as Glivenko-Cantelli theory for associated empirical processes of U-statistic structure. Illustrative applications are treated, including nonparametric and robust location and spread estimation, nonparametric analysis of linear models, nonparametric regression, and robust parametric scale estimation for exponential distributions, equivalently tail index estimation for Pareto distributions.

Keywords and phrases: Generalized L-statistics; Robust estimation

---

NOTE: This is the final version of this paper, appearing in *Proc. IISA 98 International Conference*, N. Balakrishnan (ed.), Gordon & Breach, 2000. Actual chapter number, section numbers, equation numbers, etc., differ in the published paper, and this footnote (written February, 2000) is absent.

1.1 INTRODUCTION

The notion of generalized L-statistics (GL-statistics) unifies the simpler classes of L- and U-statistics while maintaining a nice level of mathematical tractability. In applications, the notion leads to formulation of highly competitive estimators in both nonparametric and robust parametric estimation contexts. Here we review the theory and applications of GL-statistics and illustrate through several examples an interesting and potent principle, that "smoothing" followed by "medianing" yields a very favorable combination of efficiency and robustness.

Initially we consider the setting of a sample of i.i.d. real-valued observations $X_1, \ldots, X_n$ having cdf F. Denote the ordered observations by $X_{n1} \leq \cdots \leq X_{nn}$. We ask

> *What common or unifying feature is shared by the sample mean, sample variance, sample median, 5% trimmed mean, Hodges-Lehmann location estimator (i.e., median of pairwise averages $(X_i + X_j)/2$), median of three-way averages $(X_i + X_j + X_k)/3$), Theil's nonparametric regression slope estimator (i.e., median of pairwise slopes $(Y_i - Y_j)/(X_i - X_j)$), and median of absolute differences $|X_i - X_j|$ $(i \neq j)$?*

Note that among these the sample mean, sample median, and 5% trimmed mean are *L-statistics*, i.e., linear functions of order statistics given by $\sum_{i=1}^n c_{ni} X_{ni}$ for some choice of constants c_{ni}. Also, the sample mean and sample variance are *U-statistics*: i.e., for particular choices of real-valued "*kernel*" $h(x_1, \ldots, x_m)$ defined on $\mathbf{R}^m$, they can be represented in the form $n_{(m)}^{-1} \sum h(X_{i_1}, \ldots, X_{i_m})$, where the sum is over all $n_{(m)} = n(n-1)\cdots(n-m+1)$ m-tuples $(i_1, \ldots, i_m)$ of *distinct* indices from $\{1, \ldots, m\}$. Finally, the Hodges-Lehmann location estimator can be represented as an *R-statistic*, i.e., a function of the ranks of the X_i's. (General background on L-, U-, and R-statistics may be found in Huber (1981) and Serfling (1980).) The remainder of the above statistics, however, are neither L- nor U- nor R-statistics, nor do they fall within any other traditional class of statistics.

1.1.1 A Unifying Structure

We can, however, draw together *all* of the above statistics into a *single coherent class*, as follows. Consider again a kernel $h(x_1, \ldots, x_m)$ defining a U-statistic, denote the ordered values of the summands $h(X_{i_1}, \ldots, X_{i_m})$ appearing in the associated U-statistic by

$$W_{n1} \leq \cdots \leq W_{n,n_{(m)}},$$

and with these associate the class of all *linear combinations of the ordered* W_{ni}*'s*, i.e., all statistics having the form

$$\sum_{i=1}^{n_{(m)}} c_{ni} W_{ni} \tag{1.1.1}$$

for some choice of constants c_{ni}. We call statistics of form (1.1.1) *generalized L-statistics* (GL-statistics).

Note that each of the statistics considered in the above question may be expressed in this form for suitable choice of h and c_{ni}. Also, in particular, the entire class of *L-statistics* is obtained by taking kernel $h(x) = x$, and the entire class of *U-statistics* is obtained by taking $c_{ni} \equiv 1/n_{(m)}$. Moreover, interesting *new varieties* of statistic are included in this structure:

- *trimmed U-statistics* (i.e., eliminate the upper proportion α and lower proportion α W_{ni}'s and average the rest)
- *Winsorized U-statistics*
- *median of m-wise averages*, i.e., $\text{median}\{(X_{i_1} + \cdots X_{i_m})/m\}$ (which gives for $m = 1$ the usual sample median, for $m = 2$ a version of the Hodges-Lehmann location estimator, and for $m > 2$ new competitors to these estimators).

Various examples will be treated formally in §1.5.

The setting of GL-statistics may be extended in two ways. (i) The X_i's may be random elements of an *arbitrary space* as long as the kernel h is real-valued. (In the case $h(x) = x$, this reduces to requiring the X_i's to be real-valued.) (ii) In §1.2, after introducing a representation of GL-statistics in terms of *statistical functionals*, we widen this class of statistics by introducing a more general form of functional.

In order for the GL-statistic generalization to be useful in practice, the usual battery of theoretical results are needed, including asymptotic normality, strong convergence, Berry-Esséen rates, large deviation theory, sequential fixed-width confidence intervals, and jackknife and bootstrap results. These are obtained as follows. In §1.2 GL-statistics are formulated as *statistical functionals*, specifically as *L-functionals* evaluated at generalized empirical df's of *U-statistic structure*. This representation enables us in §1.3 to combine functional analysis for L-functionals with probabilistic analysis (specifically, *Glivenko-Cantelli theory*) for the generalized empirical df's, establishing a foundation for developing in §1.4 the above-mentioned theoretical results for GL-statistics. Also, in §1.4, some extensions to broader contexts are indicated. In §1.5 we examine a variety of illustrative applications in nonparametric estimation and robust parametric estimation.

1.2 BASIC FORMULATION OF GL-STATISTICS

Here we represent GL-statistics as statistical functionals. This enables a characterization of the parameter estimated by a GL-statistic as well as of the estimation error, thus providing a foundation for theoretical analysis by the method of differentiable statistical functions.

1.2.1 Representation of GL-Statistics as Statistical Functionals

Our representation of a GL-statistic as a "*differentiable statistical functional*" entails

- the use of *L-functionals* T, and
- the evaluation of such a $T(\cdot)$ at an empirical df of *U-statistic structure*.

We first review the nature of L-functionals $T(\cdot)$, then define the appropriate empirical df, and then put these together.

L-Statistics as Statistical Functionals

A functional $T(\cdot)$ defined on real-valued df's G and having the form

$$T(G) = \int_0^1 G^{-1}(t)J(t)dt \; + \; \sum_{j=1}^{d} a_j G^{-1}(p_j)$$

for some choice of function $J(\cdot)$ on [0,1], integer $d \geq 0$, values $p_j \in [0,1]$ and constants a_j, is called an *L-functional*. It represents a *weighting of the quantiles* of G, combining a *continuous* weighting of all quantiles via J with a *discrete* weighting of selected quantiles. In connection with a sample of real-valued $X_1, \ldots, X_n$ having df F, evaluation of such a $T(\cdot)$ at the usual empirical cdf

$$\widehat{F}_n(x) = n^{-1} \sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\}, \quad -\infty < x < \infty,$$

yields

$$T(\widehat{F}_n) = \sum_{i=1}^{n} \left[\int_{(i-1)/n}^{i/n} J(t)dt \right] \widehat{F}_n^{-1}(i/n) \; + \; \sum_{j=1}^{d} a_j \widehat{F}_n^{-1}(p_j),$$

which we recognize as an L-statistic because $\widehat{F}_n^{-1}(p) = X_{ni}$ for $(i-1)/n < p \leq i/n$. Thus a wide class of L-statistics is generated by evaluating various L-functionals at $\widehat{F}_n$.

Empirical CDF of U-Statistic Structure

Analogous to the above empirical df $\widehat{F}_n$ which jumps $1/n$ at the order statistics X_{ni}, we define an empirical df associated with the W_{ni}'s given above, namely the step function with jumps of size $1/n_{(m)}$:

$$\widehat{H}_n(y) = n_{(m)}^{-1} \sum \mathbf{1}\{h(X_{i_1}, \ldots, X_{i_m}) \leq y\}, \quad -\infty < y < \infty.$$

For each fixed y, $\widehat{H}_n(y)$ is a *U-statistic* as defined above. Thus, although this generalization of the usual empirical cdf has complex structure, it is of a familiar type. Note that $\widehat{H}_n$ estimates the df H_F of $h(X_1, \ldots, X_m)$:

$$E\widehat{H}_n(y) = H_F(y) = P(h(X_1, \ldots, X_m) \leq y), \quad -\infty < y < \infty.$$

For the kernel $h(x) = x$, H_F reduces to F and $\widehat{H}_n$ to $\widehat{F}_n$.

GL-Statistics as Statistical Functionals

In the same way that L-functionals evaluated at F_n yield L-statistics, we generate GL-statistics by evaluating these *same* L-functionals at the *generalized* empirical df $\widehat{H}_n$, producing

$$T(\widehat{H}_n) = \sum_{i=1}^{n_{(m)}} \left[\int_{(i-1)/n_{(m)}}^{i/n_{(m)}} J(t)dt \right] \widehat{H}_n^{-1}(i/n_{(m)}) \; + \; \sum_{j=1}^{d} a_j \widehat{H}_n^{-1}(p_j). \quad (1.2.1)$$

A wide class of linear combinations of the W_{ni}'s is thus generated. Moreover, through this representation we easily characterize the *parameter* that is estimated by a GL-statistic. Quite simply, since $\widehat{H}_n$ estimates H_F, $T(\widehat{H}_n)$ estimates

$$T(H_F) = \int_0^1 H_F^{-1}(t)J(t)dt \; + \; \sum_{j=1}^{d} a_j H_F^{-1}(p_j).$$

In the following we shall treat GL-statistics in the form (1.2.1) as well as in an extended form now to be introduced.

1.2.2 A More General Form of Functional

Let us generalize the above L-functional to:

$$T(G) = \int_0^1 q \circ T_t(G) \; J^\star(t) \; dt \; + \; \sum_{j=1}^{D} A_j \; q \circ T_{P_j}(G), \quad (1.2.2)$$

where

- for each $t \in (0,1)$, $T_t(\cdot)$ denotes a particular L-functional as defined above (with $J(\cdot)$ replaced by a function $J_t(\cdot)$, d replaced by d_t, each a_j by a_{tj}, each p_j by p_{tj})
- $q : \mathbf{R} \mapsto \mathbf{R}$ is a Borel-measurable function.

With $q(x) = x$ and $T_t(G) = G^{-1}(t)$, each t, we recover the case of simple L-functionals. Below we shall see other useful cases of $q(\cdot)$ and $T_t(\cdot)$.

Two Examples: Spread Measures of Bickel and Lehmann

Evaluation of the functional (1.2.2) at either the classical empirical df $\widehat{F}_n$ or the more general empirical df $\widehat{H}_n$ brings further statistics of interest into our scope. As examples, we mention two spread measures which Bickel and Lehmann (1979) formulated on an intuitive basis but which are best studied theoretically through reformulation as GL-statistics.

Example 1. Use (1.2.2) with $q(x) = x^2$, $T_t(G) = G^{-1}(t) - G^{-1}(1-t)$, $J^\star(t) = (1-2\beta)^{-1}$ for $\beta \le t \le 1-\beta$ and $= 0$ elsewhere, where β is chosen in $(0, 1/2)$, $D = 0$, and take $h(x) = x$ in defining $\widehat{H}_n$ (i.e., take $\widehat{F}_n$). Then the relevant GL-statistic is essentially

$$(n - 2[n\beta])^{-1} \sum_{k=[n\beta]+1}^{n-[n\beta]} (X_{nk} - X_{n,n-k+1})^2,$$

a *nonparametric* measure of *spread*. Note that in this example the more general functional $T(\cdot)$ is applied to the *classical* empirical df. □

Example 2. Use (1.2.2) with $q(x) = x^2$, $T_t(G) = G^{-1}(\frac{t+1}{2})$, $J^\star(t) = (1-\alpha-\beta)^{-1}$ for $\alpha \le t \le 1-\beta$ and $= 0$ elsewhere, where $0 < \alpha < 1/2 < 1-\beta < 1$, $D = 0$, and take $h(x_1, x_2) = x_1 - x_2$ in defining H_n. Then $T(H_n)$ yields still another nonparametric measure of spread, one which involves *both* the more general functional $T(\cdot)$ and the more general empirical df $\widehat{H}_n$. □

1.2.3 The Estimation Error

Our general goal is to study the *estimation error*,

$$T(\widehat{H}_n) - T(H_F),$$

where $T(\widehat{H}_n)$ is given by (1.2.1) using a simple L-functional, or, more generally, with $T(\cdot)$ given by a functional of form (1.2.2).

1.3 SOME FOUNDATIONAL TOOLS

We combine *functional analysis* for the functional $T(\cdot)$ with *probabilistic analysis* for the empirical cdf $\widehat{H}_n$. A convenient representation for the latter is

$$\widehat{H}_n = n_{(m)}^{-1} \sum \delta_{h(X_{i_1},\ldots,X_{i_m})},$$

where δ_y denotes the cdf placing mass 1 at the point y.

1.3.1 Differentiation Methodology

For some purposes, we require the functional $T(\cdot)$ to be *differentiable*, for which a quite basic form of differential serves very well. For an arbitrary functional $T(\cdot)$ on df's G, the *Gâteaux differential* at G_0 is defined by

$$T'(G_0; G_1 - G_0) = \frac{d}{d\lambda} T(G_0 + \lambda(G_1 - G_0))\bigg|_{0+}.$$

As is well-known (e.g., Serfling (1980)), this yields an *approximation* to $T(G_1) - T(G_0)$, when G_1 is "close" to G_0. To apply this to our object of study, the *estimation error*, we take $G_0 = H_F$ and $G_1 = \widehat{H}_n$, obtaining

$$\begin{aligned} T(\widehat{H}_n) - T(H_F) &\doteq T'(H_F; \widehat{H}_n - H_F) \\ &= T'(H_F; n_{(m)}^{-1} \sum \delta_{h(X_{i_1},\ldots,X_{i_m})} - H_F) \\ &= n_{(m)}^{-1} \sum T'(H_F; \delta_{h(X_{i_1},\ldots,X_{i_m})} - H_F), \end{aligned}$$

where in the last step *linearity* of T' in its *second* argument is assumed (to be checked for each specific functional T under consideration). Thus, for *any* functional T whose Gâteaux differential satisfies the above linearity property, the corresponding approximation to the estimation error $T(\widehat{H}_n) - T(H_F)$ has the form of a *U-statistic*, based on the "kernel"

$$T'(H_F; \delta_{h(x_1,\ldots,x_m)} - H_F). \tag{1.3.1}$$

I.e., under $\boxed{\text{linearity}}$ we have:

$$\boxed{\text{The differential approximation to the estimation error is a U-statistic.}}$$

In particular, for T an *L-functional*, and for the case that the df G_0 has density g_0, we obtain after some manipulations (Serfling (1980))

$$\begin{aligned} T'(G_0; G_1 - G_0) &= -\int_{\infty}^{\infty} [G_1(y) - G_0(y)]\ J(G_0(y))\ dy \\ &\quad + \sum_{j=1}^{d} a_j\ \frac{p_j - G_1(G_0^{-1}(p_j))}{g_0(G_0^{-1}(p_j))}. \end{aligned} \tag{1.3.2}$$

More generally, for T given by (1.2.2) with q *differentiable*, we have

$$\begin{aligned} T'(G_0; G_1 - G_0) &= -\int_0^1 q' \circ T_t(G_0)\ T_t'(G_0; G_1 - G_0)\ J^{\star}(t)\ dt \\ &\quad + \sum_{j=1}^{D} A_j\ q' \circ T_{P_j}(G)\ T'_{P_j}(G_0; G_1 - G_0), \end{aligned} \tag{1.3.3}$$

with the quantities $T_t'(G_0; G_1 - G_0)$ being of form (1.3.2). We see that the desired *linearity* of T' indeed holds, whereby we have: *for GL-statistics, the differential approximation to the estimation error is a U-statistic*. For explicit formulation of the relevant kernel given by (1.3.1), see Serfling (1984) and Janssen, Serfling, and Veraverbeke (1984). Here we simply note that the kernel in (1.3.1) has *mean* 0 and we denote its *variance* by

$$\sigma^2(T, H_F) = \mathrm{Var}(\, T'(H_F; \delta_{h(X_1,\ldots,X_m)} - H_F)\,).$$

1.3.2 The Estimation Error in the U-Empirical Process

The closeness of $T(\widehat{H}_n)$ to $T(H_F)$ is related, of course, to the closeness of $\widehat{H}_n$ to H_F. This becomes manifest in various ways. For example, to establish *asymptotic normality* of $T(\widehat{H}_n) - T(H_F)$, the relevant consideration is the behavior of the normalized difference

$$n^{1/2}[T(\widehat{H}_n) - T(H_F) - T'(H_F; \widehat{H}_n - H_F)],$$

for which a precise treatment entails the use of *rates* for the convergence of $\widehat{H}_n$ to H_F in various norms.

On the other hand, to establish the *SLLN* for $T(\widehat{H}_n)$, the relevant considerations are *continuity* rather than differentiability of $T(\cdot)$, combined with convergence of the *quantile functions* $\widehat{H}_n^{-1}$ to H_F^{-1} in various modes of convergence.

Thus the "U-empirical process" which underlies our investigation of GL-statistics becomes itself a target of investigation. The general goal is to establish for $\widehat{H}_n$ the wide collection of results already available for the classical empirical cdf $\widehat{F}_n$.

The first general result for the empirical process of U-statistic structure appears to have been developed by Silverman (1976), in work preceding the appearance of "GL-statistics" and motivated by other considerations. Indeed, treating a larger class of empirical processes, he established *weak convergence* of $n^{1/2}(\widehat{H}_n(\cdot) - H_F(\cdot))$ to a Gaussian process. In Silverman (1983), specifically for the context of GL-statistics, extension with respect to a stronger topology was obtained. One can also treat the the empirical process of U-statistic structure as a special case of "U-process" as introduced by Nolan and Pollard (1987, 1988), for which a general treatment of weak and strong convergence is provided by Arcones and Giné (1993). For a large deviation result for U-processes, see Serfling and Wang (1998).

1.3.3 Extended Glivenko-Cantelli Theory

One class of results for $\widehat{H}_n$ covers the convergence of $\widehat{H}_n$ to H_F in various modes and norms. We call this "Glivenko-Cantelli theory," in a broad sense of the term.

Results for $\|\widehat{H}_n - H_F\|_\infty$

An exponential probability inequality for $\|\widehat{H}_n - H_F\|_\infty$ was established by Helmers, Janssen, and Serfling (1988):

$$P(\,\|\widehat{H}_n - H_F\|_\infty > d\,) \le (1 + 4C[n/m]^{1/2}d)\; e^{-2[n/m]d^2}, \quad d > 0,\ n \ge m,$$

where C is a universal constant and $[\cdot]$ denotes "integer part." This is an analogue of the Dvoretzky, Kiefer, and Wolfowitz (1956) inequality for $\|\widehat{F}_n - F\|_\infty$. In fact, the latter inequality is used as a lemma in Helmers, Janssen, and Serfling (1988) to obtain an exponential bound on the *moment generating function* of $\|\widehat{H}_n - H_F\|_\infty$, thus providing a new tool even for the case $\widehat{F}_n$. As a corollary of the above probability inequality, we readily obtain

$$\|\widehat{H}_n - H_F\|_\infty \;=\; \mathrm{O}\left(\frac{\log n}{n}\right)^{1/2} \textit{almost surely},\ n \to \infty,$$

which gives the "Glivenko-Cantelli Theorem" for $\widehat{H}_n$ along with a rate of convergence. Compare the "in-probability" version,

$$\|\widehat{H}_n - H_F\|_\infty \;=\; \mathrm{O}_p(n^{-1/2}),\ n \to \infty,$$

for H_F *continuous*, proved in Serfling (1984).

The above probability inequality for $\|\widehat{H}_n - H_F\|_\infty$ also has a *multi-sample* extension, given in Helmers, Janssen, and Serfling (1988). Another variant concerns *weighted* versions of the above sup-norm, i.e.,

$$\|(\widehat{H}_n - H_F)/(w \circ H_F)\|_\infty,$$

where $w(\cdot)$ is some specified weight function. See Silverman (1983) and Helmers, Janssen, and Serfling (1988) for particular results.

Further Results

For treatment of $\|\widehat{H}_n - H_F\|_{L_p}$, see Serfling (1984), Helmers, Janssen, and Serfling (1988), and Arcones and Giné (1993), and for $\widehat{H}_n^{-1}(\cdot) - H_F^{-1}(\cdot)$, see Janssen, Serfling, and Veraverbeke (1984) and Helmers, Janssen, and Serfling (1988). Strong approximation of the U-empirical process is treated by Dehling, Denker, and Philipp (1985).

1.3.4 Oscillation Theory, Generalized Order Statistics, and Bahadur Representations

A classical *nonparametric* approach for obtaining a *confidence interval* for a *quantile parameter* $F^{-1}(p)$ is to take as endpoints of the interval a pair of *order statistics*,

$$\left(X_{n,a(n)}\ ,\ X_{n,b(n)}\right),$$

with the ranks $a(n), b(n)$ selected to achieve desired confidence. Extension to *sequential fixed-width* nonparametric C. I.'s is obtained by letting n be defined suitably as a random *stopping time* N.

A much more general and interesting class of parameters is defined by retaining the simplicity of the quantile functional,

$$T(G) = G^{-1}(p),$$

with G given by H_F based on various choices of kernel $h(x_1, \ldots, x_m)$. We have seen several examples above. For such parameters we may form nonparametric C. I.'s by taking as endpoints a suitably chosen pair of *generalized order statistics*,

$$\left(W_{n,a(n)}\ ,\ W_{n,b(n)}\right),$$

letting n be given by a stopping time N in the case of a *sequential* procedure.

Such applications are based on theoretical results for the behavior of *sequences* of the generalized order statistics, $W_{n,k(n)}$, for certain choices of rank sequence $k(n)$. A key result is a "*Bahadur-type representation*": for $0 < p < 1$, H_F twice differentiable with $H_F'(H_F^{-1}(p)) > 0$, and $k(n)$ satisfying

$$\frac{k(n)}{n_{(m)}} - p = \mathrm{o}\left(\left(\frac{\log n}{n}\right)^{1/2}\right),$$

we have that *almost surely* as $n \to \infty$

$$W_{n,k(n)} = H_F^{-1}(p) \ + \ \frac{k(n)/n_{(m)} - \widehat{H}_n(H_F^{-1}(p))}{H_F'(H_F^{-1}(p))} \ + \ \mathrm{O}(\, n^{-3/4}(\log n)^{-3/4}\,).$$

In particular, this yields for the (generalized) pth quantile $\widehat{H}_n^{-1}(p)$ a representation as *approximately a sample mean in form.*

A fundamental result on which the above result is based concerns the *oscillation behavior* of the empirical process based on $\widehat{H}_n$. Denote by

$$\omega(g;\delta) = \sup_{|s-t|\le\delta} |g(s) - g(t)|,$$

the *modulus of continuity* function for a given function g, and by

$$\alpha_n(\cdot) = n^{1/2}[\widehat{H}_n(\cdot) - H_F(\cdot)]$$

the empirical process based on $\widehat{H}_n$. Results on the rate of convergence to 0 of $\omega(\alpha_n; a_n)$ and related quantities, for sequences a_n tending to 0 at appropriate rates, are given in Silverman (1983), Janssen, Serfling, and Veraverbeke (1984), and Choudhury and Serfling (1988). In particular, the latter paper provides a broad treatment including general application to the context of sequential fixed-width nonparametric C. I.'s. The results sharpen and extend previous work of Bahadur (1966) for the case $h(x) = x$ (see also Serfling (1980)) and of Geertsema (1970) for both the cases $h(x) = x$ and $h(x_1, x_2) = \frac{1}{2}(x_1 + x_2)$. For extension to the *multi-sample* case, see Serfling (1992).

1.3.5 Estimation of the Variance of a U-Statistic

The evaluation of the Gâteaux differential of a GL-functional at $\widehat{H}_n - H_F$ was seen to be a U-statistic in form. The variance $\sigma^2(T, H_F)$ of the corresponding kernel (3) is the relevant variance parameter in the asymptotic normality of $T(\widehat{H}_n)$. Some applications require *estimation* of this variance parameter, e.g., for confidence intervals on $T(H_F)$.

General methodology for estimation of the variance of an ordinary U-statistic is available, for example, in Sen (1981). However, in the present case the kernel of our U-statistic involves *unknown parameters*. For GL-statistics which are *quantiles* of $\widehat{H}_n$, estimation of $\sigma^2(T, H_F)$ is treated in Choudhury and Serfling (1988).

1.4 GENERAL RESULTS FOR GL-STATISTICS

1.4.1 Asymptotic Normality and the LIL

Results on asymptotic normality of GL-statistics $T(\widehat{H}_n)$ are developed in Serfling (1984) and Helmers and Ruymgaart (1988) for for $T(\cdot)$ a classical L-functional with bounded scores and unbounded scores, respectively, and in Janssen, Serfling, and Veraverbeke (1984) for $T(\cdot)$ having the more general form (1.2.2). Under moderate regularity conditions, these statistics satisfy

$$n^{1/2}[T(\widehat{H}_n) - T(H_F)] \longrightarrow_d N(0, \sigma^2(T, H_F), \ n \to \infty.$$

For $T(\cdot)$ a simple L-functional, the development parallels the treatment of $T(\widehat{F}_n)$ (ordinary L-statistics) as in Serfling (1980). Briefly, put

$$\Delta_n = n^{1/2}[T(\widehat{H}_n) - T(H_F) - T'(H_F; \widehat{H}_n - H_F)]$$

and decompose this into $\Delta_n = \Delta_{n1} + \Delta_{n2}$, corresponding to the *continuous* (J-function) and *discrete* components of the functional T. Then, for Δ_{n1}, establish inequalities of the form

$$|\Delta_{n1}| \le \|W_{\widehat{H}_n, H_F}\|_A \ \ \|\widehat{H}_n - H_F\|_B, \tag{1.4.1}$$

where $A = \infty$ and $B = L_p$, or vice versa, and

$$W_{G,F}(y) = \begin{cases} \frac{K(G(y)) - K(F(y))}{G(y) - F(y)} - J(F(y)), & G(y) \ne F(y) \\ \\ 0, & G(y) = F(y), \end{cases}$$

with $K(t) = \int_0^t J(u)du$. This sets the stage for an analysis which motivates and exploits some of the *Glivenko-Cantelli* results for $\widehat{H}_n$ in §1.3. For the component Δ_{n2}, it turns out that this quantity is precisely that which is treated in the *Bahadur representation* result for $\widehat{H}_n$ as discussed in §1.3.4. For $T(\cdot)$ given by the more general functional (1.2.2), the treatment is somewhat more complicated.

For the LIL, a parallel approach works. For the *Berry-Esséen rate* for the convergence in the AN result and its use as a tool in the *bootstrap analysis* of GL-statistics, see Helmers, Janssen, and Serfling (1990).

1.4.2 The SLLN

The classical SLLN states that the sample mean converges almost surely to its expectation, a result that has fundamental and wide application in probability and statistics. Considering now the "statistical setting", we ask

In what generality does the SLLN hold?

For the generality of the class of *L-statistics*, a sharp SLLN was established by van Zwet (1980). This was extended to *GL-statistics* in Helmers, Janssen, and Serfling (1988): under moderate regularity conditions, we have

$$T(\widehat{H}_n) \longrightarrow_{a.s.} T(H_F),\ n \to \infty.$$

In some sense this is a weaker conclusion than asymptotic normality, but, since we thus need to establish it under weaker conditions, the problem can in principle be a harder one (and in fact *is*).

In the development of Helmers, Janssen, and Serfling (1988), the problem was handled by identifying and formulating the functional-analytic and probabilistic components inherent in the problem and then treating these separately. One first investigates the convergence behavior of the functional $T(\cdot)$ evaluated at a *deterministic* sequence of weakly convergent df's G_n. This leads to conditions on $T(\cdot)$ and on $\{G_n\}$, sufficient for convergence of $T(G_n)$ to a limit. Then one establishes, as an extended *Glivenko-Cantelli* property for $\widehat{H}_n$, that with probability 1 the random sequence of *empirical df's* $\{\widehat{H}_n\}$ indeed satisfies the conditions on $\{G_n\}$.

1.4.3 Large Deviation Theory

The *large deviation* problem, specialized to GL-statistics, is to evaluate the limit

$$\lim_{n\to\infty} \frac{\log P(\,|T(\widehat{H}_n) - T(H_F)| \geq d\,)}{n},$$

under appropriate conditions. For ordinary L-statistics as well as other functionals of $\widehat{F}_n$, this has been solved fairly completely in Groeneboom *et al* (1979). For extension to GL-statistics and other functionals of $\widehat{H}_n$, see Serfling and Wang (1999).

1.4.4 Further Results

Jackknife results were established for U-statistics, by Arvesen (1969) and for L-statistics by Parr and Schucany (1982). For GL-statistics of the simple form (1.2.1), jackknife results have been developed by Shao (1990). It is of interest to extend to the more general form (1.2.2). For *bootstrap* results for GL-statistics, see Helmers, Janssen, and Serfling (1990). *Multi-sample* GL-statistics are treated by Akritas (1986) and Serfling (1992). Generalizing the study of incomplete U-statistics by Blom (1976), *incomplete* GL-statistics based on the form (1.2.1) are investigated by Hössjer (1996). It is also of interest to extend to (1.2.2).

1.5 SOME APPLICATIONS

1.5.1 One-Sample Quantile Type Parameters

A general treatment of GL-statistics having the form $\widehat{H}_n^{-1}(p)$, for some choice of kernel h and $0 < p < 1$, is given by Choudhury and Serfling (1988). Some examples are as follows.

Location Estimation

For estimation of the location parameter θ of a symmetric and continuous cdf F, classical nonparametric estimators are provided by the median and by the median of pairwise averages (the Hodges-Lehmann location estimator). More generally, let us consider – as noted in Section 1 and proposed in Serfling (1984) – the median of m-wise averages:

$$\mathrm{HL}_{(m)} = \mathrm{median}\left\{\frac{X_{i_1} + \cdots X_{i_m}}{m}\right\}$$

(which for $m > 2$ gives competitors to the classical estimators). With the kernel $h(x_1, \ldots, x_m) = \frac{x_1+\ldots+x_m}{m}$, this is a *GL-statistic*: $\widehat{H}_n^{-1}(1/2)$. It estimates the *generalized quantile parameter* $H_F^{-1}(1/2)$. Besides the treatment of Choudhury and Serfling (1988) for this example, see also Choudhury (1989, 1990) and, for extension to *multivariate* X_i's, Chaudhuri (1992). In terms of *asymptotic relative efficiency* (ARE) with respect to $\overline{X}$ at the Normal distribution, and *breakdown point* (BP), the estimator $\mathrm{HL}_{(m)}$ exhibits a very favorable trade-off in comparison with other estimators, as shown in the following table.

Estimator	ARE	BP
median	.637	.500
25%-trim	.833	.250
$\mathrm{HL}_{(2)}$	.955	.293
$\mathrm{HL}_{(3)}$	.981	.206
$\mathrm{HL}_{(4)}$	.989	.160

We interpret this finding in the context of robust parametric estimation and arrive at the following principle:

> *The use of the median operation, after "smoothing" the data by taking a function of several observations at a time, over all subsets of the data, leads to a statistic which has a favorable combination of efficiency and robustness. I.e., smoothing followed by medianing yields both efficiency and robustness.*

A more general type of location estimator is given by taking a kernel of form $h(x_1, \ldots, x_m) = \sum_{i=1}^m \alpha_i x_i$ with $\sum_{i=1}^m \alpha_i = 1$. See Choudhury and Serfling (1988) for further discussion.

Spread Estimation

Among various measures of spread discussed by Bickel and Lehmann (1979) is the median of the distribution of $|X_1 - X_2|$, where X_1 and X_2 are independent r.v.'s having cdf F. This is a generalized quantile parameter, $H_F^{-1}(1/2)$, based on the kernel $h(x_1, x_2) = |x_1 - x_2|$.

More generally, as discussed in Choudhury and Serfling (1988), we might consider the class of spread measures and estimators corresponding to kernels of the form

$$h(x_1, \ldots, x_m) = \left| \sum_{i=1}^m \beta_i x_i \right|,$$

with $\sum_{i=1}^m \beta_i = 0$. This generalizes the above m-wise average form of kernel and extends an approach studied by Maritz, Wu and Staudte (1977).

Regression Slope Estimation

Consider the simple linear regression model $Y_i = \alpha + \beta X_i + \epsilon_i$, with $\{\epsilon_i\}$ i.i.d. r.v.'s independent of X_i, and X_i a sequence of *random* regressors. Let F denote the common cdf of the mutually independent pairs (X_i, Y_i), $1 \le i \le n$, and let H_F denote the cdf of $h((X_1, Y_1), (X_2, Y_2))$, where

$$h((x_1, y_1), (x_2, y_2)) = \frac{y_2 - y_1}{x_2 - x_1}.$$

For this choice of kernel, the nonparametric estimator of β given by Theil (1950), i.e., the median of the slopes $(Y_i - Y_j)/(X_i - X_j)$, is the corresponding GL-statistic based on the median functional: $\hat{\beta} = \widehat{H}_n^{-1}(1/2)$. The results of Choudhury and Serfling (1988) provide sequential nonparametric fixed-width confidence intervals for this classical estimator.

1.5.2 Two-Sample Location and Scale Problems

Location

Suppose $F^{(2)}(x) = F^{(1)}(x-\theta)$, and let F denote $(F^{(1)}, F^{(2)})$. For integer $m \geq 1$ consider the kernel

$$h(x_1^{(1)}, \ldots, x_m^{(1)}\,;\, x_1^{(2)}, \ldots, x_m^{(2)}) = \frac{(x_1^{(1)} + \ldots + x_m^{(1)}) - (x_1^{(2)} + \ldots + x_m^{(2)})}{m}.$$

Assuming $F^{(1)}$ continuous, we have that $H_F(\theta) = 1/2$, i.e., $\theta = H_F^{-1}(1/2)$, and a corresponding estimator is $\hat{\theta}_n = \widehat{H}_n^{-1}(1/2)$, where $n = (n_1, n_2)$, the vector of respective sample sizes. The case $m = 1$ is the shift estimator given by Hodges and Lehmann (1963), while the cases $m \geq 2$ represent new competing estimators. Note that under the null hypothesis $\theta = 0$ we have $H_F(0) = 1/2$, and a corresponding *test statistic* is given by $\widehat{H}_n(0)$. For the case $m = 2$, this test was proposed by Hollander (1967) (see also discussion in Randles and Wolfe (1979), pp. 96-97). See Serfling (1992) for a general development.

Scale

Suppose $F^{(2)}(x) = F^{(1)}((x-\theta)/\eta)$, for θ an unknown nuisance parameter and $\eta > 0$ the parameter of interest. With the kernel

$$h(x_1^{(1)}, x_2^{(1)}\,;\, x_1^{(2)}, x_2^{(2)}) = \frac{|x_2^{(1)} - x_1^{(1)}|}{|x_2^{(2)} - x_1^{(2)}|},$$

we have $\eta = H_F^{-1}(1/2)$, and a corresponding estimator is given by $\hat{\eta}_n = \widehat{H}_n^{-1}(1/2)$, Under the null hypothesis $\eta = 1$ we have $H_F(1) = 1/2$, and a corresponding *test statistic* is given by $\widehat{H}_n(1)$, as proposed by Lehmann (1951). See Serfling (1992) for a general development.

1.5.3 Robust ANOVA

Here we suppose that $F^{(j)}(x) = F_0(x - \Delta_j)$, $1 \leq j \leq c$, and consider estimation of a parameter of form

$$\theta = \sum_{j=1}^{c} d_j \Delta_j,$$

where $d_1, \ldots, d_c$ are specified constants and the Δ_j's are unknown. The problem of nonparametric estimation of θ in the case of a *contrast* ($\sum_1^c d_j =$

0) was initially studied and solved by Lehmann (1963), whose approach consists of expressing θ in the form of a linear combination of the differences $\Delta_i = \Delta_j$ and using nonparametric estimates of these. A rich literature has developed on this approach and its modifications. Using the framework of GL-statistics, however, a straightforward competing estimator may be formulated, based on the kernel

$$h(x_1^{(1)}; \cdots ; x_1^{(c)}) = \sum_{j=1}^{c} d_j x_1^{(j)}.$$

We suppose F_0 to be symmetric about 0, in which case we have $\theta = H_F^{-1}(1/2)$, and a natural estimator of θ is thus given by $\hat{\theta}_n = \widehat{H}_n^{-1}(1/2)$, where $n = (n_1, \ldots, n_c)$. Surprisingly, this estimator has not been investigated previously in the literature. This formulation also includes the case that θ is *not* a contrast. For testing the null hypothesis $\theta = \theta_0$, a natural test statistic is given by $\widehat{H}_n(\theta_0)$.

1.5.4 Robust Regression

Frees (1991) has introduced and investigated a wide class of estimators of β, in which a typical estimator is given by trimming the collection of ordered slopes $(Y_i - Y_j)/(X_i - X_j)$, and then taking a weighted average of the remaining slopes. Using an extended notion of generalized empirical cdf, he represents these as GL-statistics for appropriate choices of kernel.

1.5.5 Robust Estimation of Exponential Scale Parameter

Consider the problem of robust estimation of θ in the two-parameter *exponential* distribution $\mathrm{E}(\mu, \theta)$ having cdf

$$G(x) = 1 - e^{-(x-\mu)/\theta}, \quad x \geq \mu, \tag{1.5.1}$$

for $\theta > 0$ and $-\infty < \mu < \infty$, with μ an unknown "nuisance parameter". The maximum likelihood estimator of θ, $\hat{\theta}_{\mathrm{ML}} = \overline{X}_n - X_{n1}$, is *efficient*, being asymptotically normal with mean θ and variance θ^2/n, but is *nonrobust*, having BP = 0. Competing *trimmed mean* type estimators $\hat{\theta}_{\mathrm{T}}$ for various choices of trimming level β have been investigated by Kimber (1983a,b) and established to possess relatively high efficiency coupled with favorable robustness. It has been found, however, that these trimmed type estimators are outperformed by generalized median type estimators $\hat{\theta}_{\mathrm{GM}}$ based on suitable kernels. This finding illustrates again the general principle stated in §1.5.1. As a typical example, $\hat{\theta}_{\mathrm{T}}$ based on 10% upper and lower trimming has ARE = .85 and upper BP = .10, whereas $\hat{\theta}_{\mathrm{GM}}$ for a suitable kernel has ARE = .94 and upper BP = .13. For full details, see Brazauskas and Serfling (1999). Note that the above exponential scale estimation problem is equivalent, through exponential transformation of the data, to that of tail index estimation in a two-parameter Pareto model.

Acknowledgment

This research was supported by NSF Grant DMS-9705209.

References

Akritas, M. G. (1986). Empirical processes associated with V-statistics and a class of estimators under random sampling, *Annals of Statistics* **14**, 619–637.

Arcones, M. A. and Giné, E. (1993). Limit theorems for U-processes, *Annals of Probability* **21**, 1494–1542.

Arvesen, J. N. (1969). Jackknifing U-statistics, *Annals of Mathematical Statistics* **40**, 2076–2100.

Bahadur, R. R. (1966). A note on quantiles in large samples, *Annals of Mathematical Statistics* **37**, 577–580.

Bickel, P. J. and Lehmann, E. L. (1979). Descriptive statistics for nonparametric models, IV. Spread, In *Contributions to Statistics*, Hájek Memorial Volume (J. Jurecková, ed.), 33–40, Academia, Prague.

Blom, G. (1976). Some properties of incomplete U-statistics, *Biometrika* **63**, 573–580.

Brazauskas, V. and Serfling, R. (1999). Robust estimation of tail parameters for two-parameter Pareto and exponential models via generalized quantile statistics, preprint.

Chaudhuri, P. (1992). Multivariate location estimation using extension of R-estimates through U-statistics type approach, *Annals of Statistics*, **20**, 897–916.

Choudhury, J. (1989). Selecting the median of a special subset of $X + \cdots + X$, *Sesquicentennial Proceedings of American Statistical Association, Section on Statistical Computing*, 212–216.

Choudhury, J. (1990). Sequential confidence intervals based on generalized Hodges-Lehmann location estimators and related statistics, *Communications in Statistics*, **19**, 287–303.

Choudhury, J. and Serfling, R. J. (1988). Generalized order statistics, Bahadur representations, and sequential nonparametric fixed-width confidence intervals, *Journal of Statistical Planning and Inference*, **19**, 269–282.

Dehling, D., Denker, M. and Philipp, W. (1987). The almost sure invariance principle for the empirical process of U-statistic structure, *Annales de l'Institut Henri Poincare, Section B, Calcul des Probabilites et Statistique*, **23**, 121–134.

Dvoretzky, A., Kiefer, J. and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distributon function and of the classical multinomial estimator, *Annals of Mathematical Statistics*, **27**, 642–669.

Frees, E. W. (1991). Trimmed slope estimates for simple linear regression, *Journal of Statistical Planning and Inference*, **27**, 203–221.

Geertsema, J. C. (1970). Sequential confidence intervals based on rank tests, *Annals of Mathematical Statistics*, **41**, 1016–1026.

Groeneboom, P., Oosterhof, J. and Ruymgaart, F. H. (1979). Large deviation theorems for empirical probability measures, *Annals of Probability*, **7**, 553–586.

Helmers, R., Janssen, P. and Serfling, R. (1988). Glivenko-Cantelli properties of some generalized empirical df's and strong convergence of generalized L-statistics, *Probability Theory and Related Fields*, **79**, 75–93.

Helmers, R., Janssen, P. and Serfling, R. (1990). Berry-Esséen and bootstrap results for generalized L-statistics, *Scandinavian Journal of Statistics*, **17**, 65–77.

Helmers, R. and Ruymgaart, F. H. (1988). Asymptotic normality of generalized L-statistics with unbounded scores. *Journal of Statistical Planning and Inference*, **19**, 43–53.

Hodges, Jr., J. L. and Lehmann, E. L. (1963). Estimates of location based on rank tests, *Annals of Mathematical Statistics*, **34**, 598–611.

Hollander, M. (1967). Asymptotic efficiency of two nonparametric competitors of Wilcoxon's two sample test, *Journal of American Statistical Association*, **62**, 939–949.

Hössjer, O. (1996). Incomplete generalixed L-statistics, it Annals of Statistics **24**, 2631–2654.

Huber, Peter J. (1981). *Robust Statistics*, John Wiley & Sons, New York.

Janssen, P., Serfling, R. and Veraverbeke, N. (1984). Asymptotic normality for a general class of statistical functions and applications to measures of spread, *Annals of Statistics*, **12**, 1369–1379.

Kimber, A. C. (1983a). Trimming in gamma samples, *Applied Statistics* **32**, 7–14.

Kimber, A. C. (1983b). Comparison of some robust estimators of scale in gamma samples with known shape, *Journal of Statistical Computation and Simulation* **18**, 273–286.

Lehmann, E. L. (1951). Consistency and unbiasedness of certain nonparametric tests, *Annals of Mathematical Statistics*, **22**, 165–179.

Lehmann, E. L. (1963). Robust estimation in analysis of variance, *Annals of Mathematical Statistics*, **34**, 957–966.

Maritz, J. S., Wu, M. and Staudte, Jr., R. G. (1977). A location estimator based on a U-statistic, *Annals of Statistics*, **5**, 779–786.

Nolan, D. and Pollard, D. (1987). U-processes: rates of convergence, *Annals of Statistics*, **15**, 780–789.

Nolan, D. and Pollard, D. (1988). *Annals of Probability*, **16**, 1291–1298.

Parr, W. C. and Schucany, W. R. (1982). Jackknifing L-statistics with smooth weight functions, *Journal of American Statistical Association*, **27**, 639–646.

Randles, R. H. and Wolfe, D. A. (1979). *Introduction to the Theory of Nonparametric Statistics*, John Wiley & Sons, New York.

Sen, P. K. (1981). *Sequential Nonparametrics*, John Wiley & Sons, New York.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons, New York.

Serfling, R. J. (1984). Generalized L-, M- and R-statistics, *Annals of Statistics*, **12**, 76–86.

Serfling, R. J. (1992). Nonparametric confidence intervals for generalized quantile parameters in multi-sample contexts, In *Nonparametric Statistics and Related Topics* (A. K. Md. E. Saleh, ed.), 121–139, Elsevier Science Publishers B.V.

Serfling, R. and Wang, W. (1999). Large deviation results for U- and V-statistics and U- and V-empiricals, preprint.

Serfling, R. and Wang, W. (1998). A large deviation theorem for U-processes, preprint.

Shao, J. (1990). Jackknife variance estimators for generalized L-statistics, *Statistics and Probability Letters*, **11**, 27–32.

Silverman, B. (1976). Limit theorems for dissociated random variables, *Advances in Applied Probability*, **8**, 806–819.

Silverman, B. (1983). Convergence of a class of empirical distribution functions of dependent random variables, *Annals of Probability*, **11**, 745–751.

Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis, III, *Proceedings Kon. Ned. Akad. v. Wetensch. A*, **53**, 1397–1412.

van Zwet, W. (1980). A strong law for linear functions of order statistics, *Annals of Probability*, **8**, 986–990.