Robust estimations of semiparametric models: Moments

Tuban Lee

This manuscript was compiled on July 9, 2023

Descriptive statistics for parametric models currently rely heavily on the accuracy of distributional assumptions. Here, leveraging the structures of parametric distributions and their central moment kernel distributions, a class of estimators, consistent simultanously for both a semiparametric distribution and a distinct parametric distribution, is proposed. These efficient estimators are robust to both gross errors and departures from parametric assumptions, making them ideal for estimating the mean and central moments of common unimodal distributions. This article also illuminates the understanding of the common nature of probability distributions and the measures of them.

The potential biases of robust location estimators in estimating the population mean have been noticed for more than two centuries (1), with numerous significant attempts made to address them. In calculating a robust estimator, the procedure of identifying and downweighting extreme values inherently necessitates the formulation of distributional assumptions. Previously, it was demonstrated that, due to the presence of infinite-dimensional nuisance shape parameters, the semiparametric approach struggles to consistently address distributions with shapes more intricate than γ-symmetry. Newcomb (1886) provided the first modern approach to robust parametric estimation by developing a class of estimators that gives "less weight to the more discordant observations" (2). In 1964, Huber (3) used the minimax procedure to obtain M-estimator for the contaminated normal distribution, which has played a pre-eminent role in the later development of robust statistics. However, as previously demonstrated, under growing asymmetric departures from normality, the bias of the Huber M-estimator increases rapidly. This is a common issue in parametric robust statistics. For example, He and Fung (1999) constructed (4) a robust M-estimator for the two-parameter Weibull distribution, from which the mean and central moments can be calculated. Nonetheless, it is inadequate for other parametric distributions, e.g., the gamma, Perato, lognormal, and the generalized Gaussian distributions (SI Dataset S1). Another interesting approach is based on L-estimators, such as percentile estimators. For examples of percentile estimators for the Weibull distribution, the reader is referred to the works of Menon (1963) (5), Dubey (1967) (6), Marks (2005) (7), and Boudt, Caliskan, and Croux (2011) (8). At the outset of the study of percentile estimators, it was known that they arithmetically utilize the invariant structures of parametric distributions (5, 6). An estimator is classified as an I-statistic if it asymptotically satisfies $\mathrm{I}(\mathrm{LE}_1, \ldots, \mathrm{LE}_l) = (\theta_1, \ldots, \theta_q)$ for the distribution it is consistent, where LEs are calculated with the use of LU-statistics (defined in Subsection B), I is defined using arithmetic operations and constants but may also incorporate transcendental functions and quantile functions, and θs are the population parameters it estimates. In this article, two subclasses of I-statistics are introduced, recombined I-statistics and quantile I-statistics. Based on LU-statistics, I-statistics are naturally robust. Compared to probability density functions (pdfs) and cumulative distribution functions (cdfs), the quantile functions of many parametric distributions are more elegant. Since the expectation of an L-estimator can be expressed as an integral of the quantile function, I-statistics are often analytically obtainable. However, it is observed that even when the sample follows a gamma distribution, which belongs to the same larger family as the Weibull model, the generalized gamma distribution, a misassumption can still lead to substantial biases in Marks percentile estimator for the Weibull distribution (7) (SI Dataset S1).

On the other hand, while robust estimation of scale has also been intensively studied with established methods (9, 10), the development of robust measures of asymmetry and kurtosis lags behind, despite the availability of several approaches (11–15). The purpose of this paper is to demonstrate that, in light of previous works, the estimation of central moments can be transformed into a location estimation problem by using U-statistics, the central moment kernel distributions possess desirable properties, and by utilizing the invariant structures of unimodal distributions, a suite of robust estimators can be constructed whose biases are typically smaller than the variances (as seen in Table 1 for $n = 4096$).

A. Robust Estimations of the Central Moments. In 1976, Bickel and Lehmann (9), in their third paper of the landmark series *Descriptive Statistics for Nonparametric Models*, generalized nearly all robust scale estimators of that time as measures of the dispersion of a symmetric distribution around its center of symmetry. In 1979, the same series, they (10) proposed a class of estimators referred to as measures of spread, which consider the pairwise differences of a random variable, irrespective of its symmetry, throughout its distribution, rather than focusing on dispersion relative to a fixed point. While they had already considered one version of the trimmed standard deviation, which is essentially a trimmed second raw moment,

Significance Statement

Bias, variance, and contamination are the three main errors in statistics. Consistent robust estimation is unattainable without parametric assumptions. In this article, invariant moments are proposed as a means of achieving near-consistent and robust estimations of moments, even in scenarios where moderate violations of distributional assumptions occur, while the variances are sometimes smaller than those of the sample moments.

T.L. designed research, performed research, analyzed data, and wrote the paper.

The author declares no competing interest.

[1] To whom correspondence should be addressed. E-mail: tl@biomathematics.org

in the third paper of that series (9); in the final section of the fourth paper (10), they explored another two versions of the trimmed standard deviation based on symmetric differences and pairwise differences, the latter is modified here for comparison,

$$\left[\binom{n}{2}\left(1-\epsilon_{\mathbf{0}}-\gamma\epsilon_{\mathbf{0}}\right)\right]^{-\frac{1}{2}}\left[\sum_{i=\binom{n}{2}\gamma\epsilon_{\mathbf{0}}}^{\binom{n}{2}(1-\epsilon_{\mathbf{0}})}\left(X-X'\right)_{i}^{2}\right]^{\frac{1}{2}},$$

where $(X-X')_1 \leq \ldots \leq (X-X')_{\binom{n}{2}}$ are the order statistics of the pairwise differences, $X_{\mathbf{i}}-X_{\mathbf{j}}$, $\mathbf{i}<\mathbf{j}$, provided that $\binom{n}{2}\gamma\epsilon_{\mathbf{0}} \in \mathbb{N}$ and $\binom{n}{2}(1-\epsilon_{\mathbf{0}}) \in \mathbb{N}$. They showed that, when $\epsilon_{\mathbf{0}}=0$, the result obtained using [**??**] is equal to $\sqrt{2}$ times the sample standard deviation. The paper ended with, "We do not know a fortiori which of the measures is preferable and leave these interesting questions open."

Two examples of the impacts of that series are as follows. Oja (1981, 1983) (16, 17) provided a more comprehensive and generalized examination of these concepts, and integrated the measures of location, dispersion, and spread as proposed by Bickel and Lehmann (9, 10, 18), along with van Zwet's convex transformation order of skewness and kurtosis (1964) (19) for univariate and multivariate distributions, resulting a greater degree of generality and a broader perspective on these statistical constructs. Rousseeuw and Croux proposed a popular efficient scale estimator based on separate medians of pairwise differences taken over $\mathbf{i}$ and $\mathbf{j}$ (20) in 1993. However the importance of tackling the symmetry assumption has been greatly underestimated, as will be discussed later.

To address their open question (10), the nomenclature used in this paper is introduced as follows:

Nomenclature. Given a robust estimator, $\hat{\theta}$, which has an adjustable breakdown point, ϵ, that can approach zero asymptotically, the name of $\hat{\theta}$ comprises two parts: the first part denotes the type of estimator, and the second part represents the population parameter θ, such that $\hat{\theta} \to \theta$ as $\epsilon \to 0$. The abbreviation of the estimator combines the initial letters of the first part and the second part. If the estimator is symmetric, the upper asymptotic breakdown point, ϵ, is indicated in the subscript of the abbreviation of the estimator, with the exception of the median. For an asymmetric estimator based on quantile average, the associated γ follows ϵ.

In RESM I, it was shown that the bias of a robust estimator with an adjustable breakdown point is often monotonic with respect to the breakdown point in a semiparametric distribution. Naturally, the estimator's name should reflect the population parameter that it approaches as $\epsilon \to 0$. If multiplying all pseudo-samples by a factor of $\frac{1}{\sqrt{2}}$, then [**??**] is the trimmed standard deviation adhering to this nomenclature, since $\psi_2(x_1,x_2)=\frac{1}{2}(x_1-x_2)^2$ is the kernel function of the unbiased estimation of the second central moment by using U-statistic (21). This definition should be preferable, not only because it is the square root of a trimmed U-statistic, which is closely related to the minimum-variance unbiased estimator (MVUE), but also because the second γ-orderliness of the second central moment kernel distribution is ensured by the next exciting theorem.

Theorem A.1. *The second central moment kernel distribution generated from any unimodal distribution is second γ-ordered, provided that $\gamma \geq 0$.*

Proof. In 1954, Hodges and Lehmann established that if X and Y are independently drawn from the same unimodal distribution, $X-Y$ will be a symmetric unimodal distribution peaking at zero (22). Given the constraint in the pairwise differences that $X_{\mathbf{i}}<X_{\mathbf{j}}$, $\mathbf{i}<\mathbf{j}$, it directly follows from Theorem 1 in (22) that the pairwise difference distribution (Ξ_Δ) generated from any unimodal distribution is always monotonic increasing with a mode at zero. Since $X-X'$ is a negative variable that is monotonically increasing, applying the squaring transformation, the relationship between the original variable $X-X'$ and its squared counterpart $(X-X')^2$ can be represented as follows: $X-X'<Y-Y' \implies (X-X')^2>(Y-Y')^2$. In other words, as the negative values of $X-X'$ become larger in magnitude (more negative), their squared values $(X-X')^2$ become larger as well, but in a monotonically decreasing manner with a mode at zero. Further multiplication by $\frac{1}{2}$ also does not change the monotonicity and mode, since the mode is zero. Therefore, the transformed pdf becomes monotonically decreasing with a mode at zero. In RESM I, it was proven that a right-skewed distribution with a monotonic decreasing pdf is always second γ-ordered, which gives the desired result. □

In RESM I, it was shown that any γ-symmetric distribution is νth γ-U-ordered, suggesting that νth γ-U-orderliness does not require unimodality, e.g., a symmetric bimodal distribution is also νth U-ordered. In the SI Text of RESM I, an analysis of the Weibull distribution showed that unimodality does not assure orderliness. Theorem A.1 uncovers a profound relationship between unimodality, monotonicity, and second γ-orderliness, which is sufficient for γ-trimming inequality and γ-orderliness.

In 1928, Fisher constructed $\mathbf{k}$-statistics as unbiased estimators of cumulants (23). Halmos (1946) proved that a functional θ admits an unbiased estimator if and only if it is a regular statistical functional of degree $\mathbf{k}$ and showed a relation of symmetry, unbiasness and minimum variance (24). Hoeffding, in 1948, generalized U-statistics (25) which enable the derivation of a minimum-variance unbiased estimator from each unbiased estimator of an estimable parameter. In 1984, Serfling pointed out the speciality of Hodges-Lehmann estimator, which is neither a simple L-statistic nor a U-statistic, and considered the generalized L-statistics and trimmed U-statistics (26). Given a kernel function $h_{\mathbf{k}}$ which is a symmetric function of $\mathbf{k}$ variables, the LU-statistic is defined as:

$$LU_{h_{\mathbf{k}},\mathbf{k},k,\epsilon,\gamma,n} := LL_{k,\epsilon_{\mathbf{0}},\gamma,n}\left(\text{sort}\left(\left(h_{\mathbf{k}}\left(X_{N_1},\ldots,X_{N_{\mathbf{k}}}\right)\right)_{N=1}^{\binom{n}{\mathbf{k}}}\right)\right),$$

where $\epsilon = 1-(1-\epsilon_{\mathbf{0}})^{\frac{1}{\mathbf{k}}}$ (proven in Subsection F), $X_{N_1},\ldots,X_{N_{\mathbf{k}}}$ are the n choose $\mathbf{k}$ elements from the sample, $LL_{k,\epsilon_{\mathbf{0}},\gamma,n}(Y)$ denotes the LL-statistic with the sorted sequence $\text{sort}\left(\left(h_{\mathbf{k}}\left(X_{N_1},\ldots,X_{N_{\mathbf{k}}}\right)\right)_{N=1}^{\binom{n}{\mathbf{k}}}\right)$ serving as an input. In the context of Serfling's work, the term 'trimmed U-statistic' is used when $LL_{k,\epsilon_{\mathbf{0}},\gamma,n}$ is $\mathrm{TM}_{\epsilon_{\mathbf{0}},\gamma,n}$ (26).

In 1997, Heffernan (21) obtained an unbiased estimator of the $\mathbf{k}$th central moment by using U-statistics and demonstrated that it is the minimum variance unbiased estimator for distributions with the finite first $\mathbf{k}$ moments. The weighted Hodges-Lehmann $\mathbf{k}$th central moment ($2 \leq \mathbf{k} \leq n$) is thus defined as,

$$\mathrm{WHL}\mathbf{k}m_{k,\epsilon,\gamma,n} := LU_{h_{\mathbf{k}}=\psi_{\mathbf{k}},\mathbf{k},k,\epsilon,\gamma,n},$$

where $\mathrm{WHLM}_{k,\epsilon_{\mathbf{0}},\gamma,n}$ is used as the $LL_{k,\epsilon_{\mathbf{0}},\gamma,n}$ in LU, $\psi_{\mathbf{k}}(x_1,\ldots,x_{\mathbf{k}}) = \sum_{j=0}^{\mathbf{k}-2}(-1)^j\left(\frac{1}{\mathbf{k}-j}\right)\sum\left(x_{i_1}^{\mathbf{k}-j}x_{i_2}\ldots x_{i_{j+1}}\right) + (-1)^{\mathbf{k}-1}(\mathbf{k}-1)x_1\ldots x_{\mathbf{k}}$, the second summation is over $i_1,\ldots,i_{j+1} = 1$ to $\mathbf{k}$ with $i_1 \neq i_2 \neq \ldots \neq i_{j+1}$ and $i_2 < i_3 < \ldots < i_{j+1}$ (21). Despite the complexity, the following theorem offers an approach to infer the general structure of such kernel distributions.

Theorem A.2. *Define a set T comprising all pairs $(\psi_{\mathbf{k}}(\mathbf{v}), f_{X,\ldots,X}(\mathbf{v}))$ such that $\psi_{\mathbf{k}}(\mathbf{v}) = \psi_{\mathbf{k}}(Q(p_1),\ldots,Q(p_{\mathbf{k}}))$ with $Q(p_1) < \ldots < Q(p_{\mathbf{k}})$ and $f_{X,\ldots,X}(\mathbf{v}) = \mathbf{k}!f(Q(p_1))\ldots f(Q(p_{\mathbf{k}}))$ is the probability density of the $\mathbf{k}$-tuple, $\mathbf{v} = (Q(p_1),\ldots,Q(p_{\mathbf{k}}))$ (a formula drawn after a modification of the Jacobian density theorem). T_Δ is a subset of T, consisting all those pairs for which the corresponding $\mathbf{k}$-tuples satisfy that $Q(p_1) - Q(p_{\mathbf{k}}) = \Delta$. The component quasi-distribution, denoted by ξ_Δ, has a quasi-pdf $f_{\xi_\Delta}(\bar{\Delta}) = \sum_{\substack{(\psi_{\mathbf{k}}(\mathbf{v}),f_{X,\ldots,X}(\mathbf{v}))\in T_\Delta \\ \bar{\Delta}=\psi_{\mathbf{k}}(\mathbf{v})}} f_{X,\ldots,X}(\mathbf{v})$, i.e., sum over all $f_{X,\ldots,X}(\mathbf{v})$ such that the pair $(\psi_{\mathbf{k}}(\mathbf{v}), f_{X,\ldots,X}(\mathbf{v}))$ is in the set T_Δ and the first element of the pair, $\psi_{\mathbf{k}}(\mathbf{v})$, is equal to $\bar{\Delta}$. The $\mathbf{k}$th, where $\mathbf{k} > 2$, central moment kernel distribution, labeled $\Xi_{\mathbf{k}}$, can be seen as a quasi-mixture distribution comprising an infinite number of component quasi-distributions, ξ_Δs, each corresponding to a different value of Δ, which ranges from $Q(0) - Q(1)$ to 0. Each component quasi-distribution has a support of $\left(-\binom{\mathbf{k}}{\frac{3+(-1)^{\mathbf{k}}}{2}}^{-1}(-\Delta)^{\mathbf{k}}, \frac{1}{\mathbf{k}}(-\Delta)^{\mathbf{k}}\right)$.*

Proof. The support of ξ_Δ is the extrema of the function $\psi_{\mathbf{k}}(Q(p_1),\cdots,Q(p_{\mathbf{k}}))$ subjected to the constraints, $Q(p_1) < \cdots < Q(p_{\mathbf{k}})$ and $\Delta = Q(p_1) - Q(p_{\mathbf{k}})$. Using the Lagrange multiplier, the only critical point can be determined at $Q(p_1) = \cdots = Q(p_{\mathbf{k}}) = 0$, where $\psi_{\mathbf{k}} = 0$. Other candidates are within the boundaries, i.e., $\psi_{\mathbf{k}}(x_1 = Q(p_1), x_2 = Q(p_{\mathbf{k}}),\cdots,x_{\mathbf{k}} = Q(p_{\mathbf{k}}))$, $\cdots$, $\psi_{\mathbf{k}}(x_1 = Q(p_1),\cdots,x_i = Q(p_1), x_{i+1} = Q(p_{\mathbf{k}}),\cdots,x_{\mathbf{k}} = Q(p_{\mathbf{k}}))$, $\cdots$, $\psi_{\mathbf{k}}(x_1 = Q(p_1),\cdots,x_{\mathbf{k}-1} = Q(p_1), x_{\mathbf{k}} = Q(p_{\mathbf{k}}))$. $\psi_{\mathbf{k}}(x_1 = Q(p_1),\cdots,x_i = Q(p_1), x_{i+1} = Q(p_{\mathbf{k}}),\cdots,x_{\mathbf{k}} = Q(p_{\mathbf{k}}))$ can be divided into $\mathbf{k}$ groups. The gth group has the common factor $(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}$, if $1 \le g \le \mathbf{k}-1$ and the final $\mathbf{k}$th group is the term $(-1)^{\mathbf{k}-1}(\mathbf{k}-1)Q(p_1)^iQ(p_{\mathbf{k}})^{\mathbf{k}-i}$. If $\frac{\mathbf{k}+1-i}{2} \le j \le \frac{\mathbf{k}-1}{2}$ and $j+1 \le g \le \mathbf{k}-j$, the gth group has $i\binom{i-1}{g-j-1}\binom{\mathbf{k}-i}{j}$ terms having the form $(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}Q(p_1)^{\mathbf{k}-j}Q(p_{\mathbf{k}})^j$. If $\frac{\mathbf{k}+1-i}{2} \le j \le \frac{\mathbf{k}-1}{2}$ and $\mathbf{k}-j+1 \le g \le i+j$, the gth group has $i\binom{i-1}{g-j-1}\binom{\mathbf{k}-i}{j} + (\mathbf{k}-i)\binom{\mathbf{k}-i-1}{j-\mathbf{k}+g-1}\binom{i}{\mathbf{k}-j}$ terms having the form $(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}Q(p_1)^{\mathbf{k}-j}Q(p_{\mathbf{k}})^j$. If $0 \le j < \frac{\mathbf{k}+1-i}{2}$ and $j+1 \le g \le i+j$, the gth group has $i\binom{i-1}{g-j-1}\binom{\mathbf{k}-i}{j}$ terms having the form $(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}Q(p_1)^{\mathbf{k}-j}Q(p_{\mathbf{k}})^j$. If $\frac{\mathbf{k}}{2} \le j \le \mathbf{k}$ and $\mathbf{k}-j+1 \le g \le j$, the gth group has $(\mathbf{k}-i)\binom{\mathbf{k}-i-1}{j-\mathbf{k}+g-1}\binom{i}{\mathbf{k}-j}$ terms having the form $(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}Q(p_1)^{\mathbf{k}-j}Q(p_{\mathbf{k}})^j$. If $\frac{\mathbf{k}}{2} \le j \le \mathbf{k}$ and $j+1 \le g \le j+i < \mathbf{k}$, the gth group has $i\binom{i-1}{g-j-1}\binom{\mathbf{k}-i}{j} + (\mathbf{k}-i)\binom{\mathbf{k}-i-1}{j-\mathbf{k}+g-1}\binom{i}{\mathbf{k}-j}$ terms having the form $(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}Q(p_1)^{\mathbf{k}-j}Q(p_{\mathbf{k}})^j$. So, if $i+j = \mathbf{k}$, $\frac{\mathbf{k}}{2} \le j \le \mathbf{k}$, $0 \le i \le \frac{\mathbf{k}}{2}$, the summed coefficient of $Q(p_1)^iQ(p_{\mathbf{k}})^{\mathbf{k}-i}$ is $(-1)^{\mathbf{k}-1}(\mathbf{k}-1) + \sum_{g=i+1}^{\mathbf{k}-1}(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}(\mathbf{k}-i)\binom{\mathbf{k}-i-1}{g-i-1} + \sum_{g=\mathbf{k}-i+1}^{\mathbf{k}-1}(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}i\binom{i-1}{g-\mathbf{k}+i-1} = (-1)^{\mathbf{k}-1}(\mathbf{k}-1) + (-1)^{\mathbf{k}+1} + (\mathbf{k}-i)(-1)^{\mathbf{k}} + (-1)^{\mathbf{k}}(i-1) = (-1)^{\mathbf{k}+1}$. The summation identities are $\sum_{g=i+1}^{\mathbf{k}-1}(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}(\mathbf{k}-i)\binom{\mathbf{k}-i-1}{g-i-1} = (\mathbf{k}-i)\int_0^1\sum_{g=i+1}^{\mathbf{k}-1}(-1)^{g+1}\binom{\mathbf{k}-i-1}{g-i-1}t^{\mathbf{k}-g}dt = (\mathbf{k}-i)\int_0^1\left((-1)^i(t-1)^{\mathbf{k}-i-1} - (-1)^{\mathbf{k}+1}\right)dt = (\mathbf{k}-i)\left(\frac{(-1)^{\mathbf{k}}}{i-\mathbf{k}} + (-1)^{\mathbf{k}}\right) = (-1)^{\mathbf{k}+1} + (\mathbf{k}-i)(-1)^{\mathbf{k}}$ and $\sum_{g=\mathbf{k}-i+1}^{\mathbf{k}-1}(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}i\binom{i-1}{g-\mathbf{k}+i-1} = \int_0^1\sum_{g=\mathbf{k}-i+1}^{\mathbf{k}-1}(-1)^{g+1}i\binom{i-1}{g-\mathbf{k}+i-1}t^{\mathbf{k}-g}dt = \int_0^1\left(i(-1)^{\mathbf{k}-i}(t-1)^{i-1} - i(-1)^{\mathbf{k}+1}\right)dt = (-1)^{\mathbf{k}}(i-1)$. If $0 \le j < \frac{\mathbf{k}+1-i}{2}$ and $i = \mathbf{k}$, $\psi_{\mathbf{k}} = 0$. If $\frac{\mathbf{k}+1-i}{2} \le j \le \frac{\mathbf{k}-1}{2}$ and $\frac{\mathbf{k}+1}{2} \le i \le \mathbf{k}-1$, the summed coefficient of $Q(p_1)^iQ(p_{\mathbf{k}})^{\mathbf{k}-i}$ is $(-1)^{\mathbf{k}-1}(\mathbf{k}-1) + \sum_{g=\mathbf{k}-i+1}^{\mathbf{k}-1}(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}i\binom{i-1}{g-\mathbf{k}+i-1} + \sum_{g=i+1}^{\mathbf{k}-1}(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}(\mathbf{k}-i)\binom{\mathbf{k}-i-1}{g-i-1}$, the same as above. If $i+j < \mathbf{k}$, since $\binom{i}{\mathbf{k}-j} = 0$, the related terms can be ignored, so, using the binomial theorem and beta function, the summed coefficient of $Q(p_1)^{k-j}Q(p_{\mathbf{k}})^j$ is $\sum_{g=j+1}^{i+j}(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}i\binom{i-1}{g-j-1}\binom{\mathbf{k}-i}{j} = i\binom{\mathbf{k}-i}{j}\int_0^1\sum_{g=j+1}^{i+j}(-1)^{g+1}\binom{i-1}{g-j-1}t^{\mathbf{k}-g}dt = \binom{\mathbf{k}-i}{j}i\int_0^1\left((-1)^j t^{\mathbf{k}-j-1}\left(\frac{t}{t-1}\right)^{1-i}\right)dt = \binom{\mathbf{k}-i}{j}i\frac{(-1)^{j+i+1}\Gamma(i)\Gamma(\mathbf{k}-j-i+1)}{\Gamma(\mathbf{k}-j+1)} = \frac{(-1)^{j+i+1}i!(\mathbf{k}-j-i)!(\mathbf{k}-i)!}{(\mathbf{k}-j)!j!(\mathbf{k}-j-i)!} = (-1)^{j+i+1}\frac{i!(\mathbf{k}-i)!}{\mathbf{k}!}\frac{\mathbf{k}!}{(\mathbf{k}-j)!j!} = \binom{\mathbf{k}}{i}^{-1}(-1)^{1+i}\binom{\mathbf{k}}{j}(-1)^j$.

According to the binomial theorem, the coefficient of $Q(p_1)^iQ(p_{\mathbf{k}})^{\mathbf{k}-i}$ in $\binom{\mathbf{k}}{i}^{-1}(-1)^{1+i}(Q(p_1) - Q(p_{\mathbf{k}}))^{\mathbf{k}}$ is $\binom{\mathbf{k}}{i}^{-1}(-1)^{1+i}\binom{\mathbf{k}}{i}(-1)^{\mathbf{k}-i} = (-1)^{\mathbf{k}+1}$, same as the above summed coefficient of $Q(p_1)^iQ(p_{\mathbf{k}})^{\mathbf{k}-i}$, if $i+j = \mathbf{k}$. If $i+j < k$, the coefficient of $Q(p_1)^{\mathbf{k}-j}Q(p_{\mathbf{k}})^j$ is $\binom{\mathbf{k}}{i}^{-1}(-1)^{1+i}\binom{\mathbf{k}}{j}(-1)^j$, same as the corresponding summed coefficient of $Q(p_1)^{\mathbf{k}-j}Q(p_{\mathbf{k}})^j$. Therefore, $\psi_{\mathbf{k}}(x_1 = Q(p_1),\ldots,x_i = Q(p_1), x_{i+1} = Q(p_{\mathbf{k}}),\ldots,x_{\mathbf{k}} = Q(p_{\mathbf{k}})) = \binom{\mathbf{k}}{i}^{-1}(-1)^{1+i}(Q(p_1) - Q(p_{\mathbf{k}}))^{\mathbf{k}}$, the maximum and minimum of $\psi_{\mathbf{k}}$ follow directly from the properties of the binomial coefficient.

□

The component quasi-distribution, ξ_Δ, is closely related to Ξ_Δ, which is the pairwise difference distribution, since $\sum_{\bar{\Delta}=-\binom{\mathbf{k}}{\frac{3+(-1)^{\mathbf{k}}}{2}}^{-1}(-\Delta)^{\mathbf{k}}}^{\frac{1}{\mathbf{k}}(-\Delta)^{\mathbf{k}}} f_{\xi_\Delta}(\bar{\Delta}) = f_{\Xi_\Delta}(\Delta)$. Recall that Theorem A.1 established that $f_{\Xi_\Delta}(\Delta)$ is monotonic increasing with a mode at zero if the original distribution is unimodal, $f_{\Xi_{-\Delta}}(-\Delta)$ is thus monotonic decreasing with a mode at zero. In general, if assuming the shape of ξ_Δ is uniform, $\Xi_{\mathbf{k}}$ is monotonic left and right around zero. The median of $\Xi_{\mathbf{k}}$ also exhibits a strong tendency to be close to zero, as it can be cast as a weighted mean of the medians of ξ_Δ. When $-\Delta$ is small, all values of ξ_Δ are close to zero, resulting in the median of ξ_Δ being close to zero as well. When $-\Delta$ is large, the median of ξ_Δ depends on its skewness, but the corresponding weight is much smaller, so even if ξ_Δ is highly skewed, the median of $\Xi_{\mathbf{k}}$ will only be slightly shifted from zero. Denote the median of $\Xi_{\mathbf{k}}$ as $m\mathbf{k}m$, for the five parametric distributions here, $|m\mathbf{k}m|$s are all $\le 0.1\sigma$ for Ξ_3 and Ξ_4, where σ is the standard deviation of $\Xi_{\mathbf{k}}$ (SI Dataset S1). Assuming $m\mathbf{k}m = 0$, for the even ordinal central moment kernel distribution, the average probability density on the left side of zero is greater than that on the right side, since

$\frac{\frac{1}{2}}{\binom{\mathbf{k}}{2}^{-1}(Q(0)-Q(1))^{\mathbf{k}}} > \frac{\frac{1}{2}}{\frac{1}{\mathbf{k}}(Q(0)-Q(1))^{\mathbf{k}}}$. This means that, on average, the inequality $f(Q(\epsilon)) \geq f(Q(1-\epsilon))$ holds. For the odd ordinal distribution, the discussion is more challenging since it is generally symmetric. Just consider Ξ_3, let $x_1 = Q(p_i)$ and $x_3 = Q(p_j)$, changing the value of x_2 from $Q(p_i)$ to $Q(p_j)$ will monotonically change the value of $\psi_3(x_1, x_2, x_3)$, since $\frac{\partial \psi_3(x_1,x_2,x_3)}{\partial x_2} = -\frac{x_1^2}{2} - x_1x_2 + 2x_1x_3 + x_2^2 - x_2x_3 - \frac{x_3^2}{2}$, $-\frac{3}{4}(x_1 - x_3)^2 \leq \frac{\partial \psi_3(x_1,x_2,x_3)}{\partial x_2} \leq -\frac{1}{2}(x_1 - x_3)^2 \leq 0$. If the original distribution is right-skewed, ξ_Δ will be left-skewed, so, for Ξ_3, the average probability density of the right side of zero will be greater than that of the left side, which means, on average, the inequality $f(Q(\epsilon)) \leq f(Q(1-\epsilon))$ holds. In all, the monotonic decreasing of the negative pairwise difference distribution guides the general shape of the $\mathbf{k}$th central moment kernel distribution, $\mathbf{k} > 2$, forcing it to be unimodal-like with the mode and median close to zero, then, the inequality $f(Q(\epsilon)) \leq f(Q(1-\epsilon))$ or $f(Q(\epsilon)) \geq f(Q(1-\epsilon))$ holds in general. If a distribution is νth γ-ordered and all of its central moment kernel distributions are also νth γ-ordered, it is called completely νth γ-ordered. Although strict complete νth γ-orderliness is difficult to prove, even if the inequality may be violated in a small range, as discussed in Subsection **??**, the mean-SWA$_\epsilon$-median inequality remains valid, in most cases, for the central moment kernel distribution.

To avoid confusion, it should be noted that the robust location estimations of the kernel distributions discussed in this paper differ from the approach taken by Joly and Lugosi (2016) (27), which is computing the median of all U-statistics from different disjoint blocks. Compared to bootstrap median U-statistics, this approach can produce two additional kinds of finite sample bias, one arises from the limited numbers of blocks, another is due to the size of the U-statistics (consider the mean of all U-statistics from different disjoint blocks, it is definitely not identical to the original U-statistic, except when the kernel is the Hodges-Lehmann kernel). Laforgue, Clemencon, and Bertail (2019)'s median of randomized U-statistics (28) is more sophisticated and can overcome the limitation of the number of blocks, but the second kind of bias remains unsolved.

B. Invariant Moments. All popular robust location estimators, such as the symmetric trimmed mean, symmetric Winsorized mean, Hodges-Lehmann estimator, Huber M-estimator, and median of means, are symmetric. As shown in RESM I, a γ-weighted Hodges-Lehmann mean ($\mathrm{WHLM}_{k,\epsilon,\gamma}$) can achieve consistency for the population mean in any γ-symmetric distribution with a finite mean. However, it falls considerably short of consistently handling other parametric distributions that are not γ-symmetric. Shifting from semiparametrics to parametrics, consider a robust estimator with a non-sample-dependent breakdown point (defined in Subsection F) which is consistent simultaneously for both a semiparametric distribution and a parametric distribution that does not belong to that semiparametric distribution, it is named with the prefix 'invariant' followed by the name of the population parameter it is consistent with. Here, the recombined I-statistic is defined as

$$\mathrm{RI}_{d,h_{\mathbf{k}},\mathbf{k}_1,\mathbf{k}_2,k_1,k_2,\epsilon=\min(\epsilon_1,\epsilon_2),\gamma_1,\gamma_2,n,LU_1,LU_2} := \lim_{c\to\infty}\left(\frac{\left(LU_{1\,h_{\mathbf{k}},\mathbf{k}_1,k_1,\epsilon_1,\gamma_1,n}+c\right)^{d+1}}{\left(LU_{2\,h_{\mathbf{k}},\mathbf{k}_2,k_2,\epsilon_2,\gamma_2,n}+c\right)^{d}} - c\right),$$

where d is the key factor for bias correction, $LU_{h_{\mathbf{k}},\mathbf{k},k,\epsilon,\gamma,n}$ is the LU-statistic, $\mathbf{k}$ is the degree of the U-statistic, k is the degree of the LL-statistic, ϵ is the upper asymptotic breakdown point of the LU-statistic. It is assumed in this series that in the subscript of an estimator, if $\mathbf{k}$, k and γ are omitted, $\mathbf{k} = 1$, $k = 1$, $\gamma = 1$ are assumed, if just one $\mathbf{k}$ is indicated, $\mathbf{k}_1 = \mathbf{k}_2$, if just one γ is indicated, $\gamma_1 = \gamma_2$, if n is omitted, only the asymptotic behavior is considered, in the absence of subscripts, no assumptions are made. The subsequent theorem shows the significance of a recombined I-statistic.

Theorem B.1. *Define the recombined mean as* $rm_{d,k_1,k_2,\epsilon=\min(\epsilon_1,\epsilon_2),\gamma_1,\gamma_2,n,WL_1,WL_2} := RI_{d,h_{\mathbf{k}}=x,\mathbf{k}_1=1,\mathbf{k}_2=1,k_1,k_2,\epsilon=\min(\epsilon_1,\epsilon_2),\gamma_1,\gamma_2,n,LU_1=WL_1,LU_2=WL_2}$. *Assuming finite means,*

$$rm_{d=\frac{\mu - WL_{1\,k_1,\epsilon_1,\gamma_1}}{WL_{1\,k_1,\epsilon_1,\gamma_1} - WL_{2\,k_2,\epsilon_2,\gamma_2}},k_1,k_2,\epsilon=\min(\epsilon_1,\epsilon_2),\gamma_1,\gamma_2,WL_1,WL_2}$$

is a consistent mean estimator for a location-scale distribution, where μ, $WL_{1\,k_1,\epsilon_1,\gamma_1}$, *and* $WL_{2\,k_2,\epsilon_2,\gamma_2}$ *are different location parameters from that location-scale distribution. If* $\gamma_1 = \gamma_2$, $WL = WHLM$, rm *is also consistent for any* γ*-symmetric distributions.*

Proof. Finding d that make $rm_{d,k_1,k_2,\epsilon=\min(\epsilon_1,\epsilon_2),\gamma_1,\gamma_2,\mathrm{WL}_1,\mathrm{WL}_2}$ a consistent mean estimator is equivalent to finding the solution of $rm_{d,k_1,k_2,\epsilon=\min(\epsilon_1,\epsilon_2),\gamma_1,\gamma_2,\mathrm{WL}_1,\mathrm{WL}_2} = \mu$. First consider the location-scale distribution. Since $rm_{d,k_1,k_2,\epsilon=\min(\epsilon_1,\epsilon_2),\gamma_1,\gamma_2,\mathrm{WL}_1,\mathrm{WL}_2} = \lim_{c\to\infty}\left(\frac{\left(\mathrm{WL}_{1\,k_1,\epsilon_1,\gamma_1}+c\right)^{d+1}}{\left(\mathrm{WL}_{2\,k_2,\epsilon_2,\gamma_2}+c\right)^{d}} - c\right) = (d+1)\,\mathrm{WL}_{1\,k_1,\epsilon_1,\gamma} - d\mathrm{WL}_{2\,k_2,\epsilon_2,\gamma} = \mu$. So, $d = \frac{\mu - \mathrm{WL}_{1\,k_1,\epsilon_1,\gamma_1}}{\mathrm{WL}_{1\,k_1,\epsilon_1,\gamma_1} - \mathrm{WL}_{2\,k_2,\epsilon_2,\gamma_2}}$. In RESM I, it was established that any $\mathrm{WL}(k,\epsilon,\gamma)$ can be expressed as $\lambda\mathrm{WL}_0(k,\epsilon,\gamma) + \mu$ for a location-scale distribution parameterized by a location parameter μ and a scale parameter λ, where $\mathrm{WL}_0(k,\epsilon,\gamma)$ is a function of $Q_0(p)$, the quantile function of a standard distribution without any shifts or scaling, according to the definition of the weighted L-statistic. The simultaneous cancellation of μ and λ in $\frac{(\lambda\mu_0+\mu)-(\lambda\mathrm{WL}_{1\,0}(k_1,\epsilon_1,\gamma_1)+\mu)}{(\lambda\mathrm{WL}_{1\,0}(k_1,\epsilon_1,\gamma_1)+\mu)-(\lambda\mathrm{WL}_{2\,0}(k_2,\epsilon_2,\gamma_2)+\mu)}$ assures that the d in rm is always a constant for a location-scale distribution. The proof of the second assertion follows directly from the coincidence property. According to RESM I, for any γ-symmetric distribution with a finite mean, $\mathrm{WHLM}_{1\,k_1,\epsilon_1,\gamma} = \mathrm{WHLM}_{2\,k_2,\epsilon_2,\gamma} = \mu$. Then $rm_{d,k_1,k_2,\epsilon_1,\epsilon_2,\gamma,\mathrm{WHLM}_1,\mathrm{WHLM}_2} = \lim_{c\to\infty}\left(\frac{(\mu+c)^{d+1}}{(\mu+c)^{d}} - c\right) = \mu$. This completes the demonstration. □

For example, the Pareto distribution has a quantile function $Q_{Par}(p) = x_m(1-p)^{-\frac{1}{\alpha}}$, where x_m is the minimum possible value that a random variable following the Pareto distribution can take, serving a scale parameter, α is a shape parameter. The mean of the Pareto distribution is given by $\frac{\alpha x_m}{\alpha-1}$. As $\mathrm{WL}(k,\epsilon,\gamma)$ can be expressed as a function of $Q(p)$, one can set the two $\mathrm{WL}_{k,\epsilon,\gamma}$s in the d value of rm as two arbitrary quantiles $Q_{Par}(p_1)$ and $Q_{Par}(p_2)$. For the Pareto distribution, $d_{Per,rm} = \frac{\mu_{Per}-Q_{Par}(p_1)}{Q_{Par}(p_1)-Q_{Par}(p_2)} = \frac{\frac{\alpha x_m}{\alpha-1}-x_m(1-p_1)^{-\frac{1}{\alpha}}}{x_m(1-p_1)^{-\frac{1}{\alpha}}-x_m(1-p_2)^{-\frac{1}{\alpha}}}$. x_m can be canceled out. Intriguingly, the quantile function of exponential distribution is $Q_{exp}(p) = \ln\left(\frac{1}{1-p}\right)\lambda$, $\lambda \geq 0$. $\mu_{exp} = \lambda$. Then, $d_{exp,rm} = \frac{\mu_{exp}-Q_{exp}(p_1)}{Q_{exp}(p_1)-Q_{exp}(p_2)} = \frac{\lambda-\ln\left(\frac{1}{1-p_1}\right)\lambda}{\ln\left(\frac{1}{1-p_1}\right)\lambda-\ln\left(\frac{1}{1-p_2}\right)\lambda} = -\frac{\ln(1-p_1)+1}{\ln(1-p_1)-\ln(1-p_2)}$. Since $\lim_{\alpha\to\infty} \frac{\frac{\alpha}{\alpha-1}-(1-p_1)^{-1/\alpha}}{(1-p_1)^{-1/\alpha}-(1-p_2)^{-1/\alpha}} = -\frac{\ln(1-p_1)+1}{\ln(1-p_1)-\ln(1-p_2)}$, $d_{Per,rm}$ approaches $d_{exp,rm}$, as $\alpha \to \infty$, regardless of the type of weighted L-statistic used. That means, for the Weibull, gamma, Pareto, lognormal and generalized Gaussian distribution, $rm_{d=\frac{\mu-\mathrm{WHLM}_{1k_1,\epsilon_1,\gamma}}{\mathrm{WHLM}_{1k_1,\epsilon_1,\gamma}-\mathrm{WHLM}_{2k_2,\epsilon_2,\gamma}},k_1,k_2,\epsilon=\min(\epsilon_1,\epsilon_2),\gamma,\mathrm{WHLM}_1,\mathrm{WHLM}_2}$ is consistent for at least one particular case, where μ, $\mathrm{WHLM}_{1k_1,\epsilon_1,\gamma}$, and $\mathrm{WHLM}_{2k_2,\epsilon_2,\gamma}$ are different location parameters from an exponential distribution. Let $\mathrm{WHLM}_{1k_1,\epsilon_1,\gamma} = \mathrm{BM}_{\nu=3,\epsilon=\frac{1}{24}}$, $\mathrm{WHLM}_{2k_2,\epsilon_2,\gamma} = m$, then $\mu = \lambda$, $m = Q\left(\frac{1}{2}\right) = \ln 2\lambda$, $\mathrm{BM}_{\nu=3,\epsilon=\frac{1}{24}} = \lambda\left(1+\ln\left(\frac{26068394603446272\sqrt[6]{\frac{7}{247}}\sqrt[3]{11}}{391^{5/6}101898752449325\sqrt{5}}\right)\right)$, the detailed formula is given in the SI Text. So, $d = \frac{\mu-\mathrm{BM}_{\nu=3,\epsilon=\frac{1}{24}}}{\mathrm{BM}_{\nu=3,\epsilon=\frac{1}{24}}-m} = \frac{\lambda-\lambda\left(1+\ln\left(\frac{26068394603446272\sqrt[6]{\frac{7}{247}}\sqrt[3]{11}}{391^{5/6}101898752449325\sqrt{5}}\right)\right)}{\lambda\left(1+\ln\left(\frac{26068394603446272\sqrt[6]{\frac{7}{247}}\sqrt[3]{11}}{391^{5/6}101898752449325\sqrt{5}}\right)\right)-\ln 2\lambda} = -\frac{\ln\left(\frac{26068394603446272\sqrt[6]{\frac{7}{247}}\sqrt[3]{11}}{391^{5/6}101898752449325\sqrt{5}}\right)}{1-\ln(2)+\ln\left(\frac{26068394603446272\sqrt[6]{\frac{7}{247}}\sqrt[3]{11}}{391^{5/6}101898752449325\sqrt{5}}\right)} \approx 0.103$. The biases of $rm_{d\approx0.103,\nu=3,\epsilon=\frac{1}{24},\mathrm{BM},m}$ for distributions with skewness between those of the exponential and symmetric distributions are tiny (SI Dataset S1). $rm_{d\approx0.103,\nu=3,\epsilon=\frac{1}{24},\mathrm{BM},m}$ exhibits excellent performance for all these common unimodal distributions (SI Dataset S1).

The recombined mean is an recombined I-statistic. Consider an I-statistic whose LEs are percentiles of a distribution obtained by plugging LU-statistics into a cumulative distribution function, I is defined with arithmetic operations, constants and quantile functions, such an estimator is classified as a quantile I-statistic. One version of the quantile I-statistic can be defined as $\mathrm{QI}_{d,h_{\mathbf{k}},\mathbf{k},k,\epsilon,\gamma,,n,LU} :=$

$$\begin{cases} \hat{Q}_{n,h_{\mathbf{k}}}\left(\left(\hat{F}_{n,h_{\mathbf{k}}}(LU)-\frac{\gamma}{1+\gamma}\right)d+\hat{F}_{n,h_{\mathbf{k}}}(LU)\right) & \hat{F}_{n,h_{\mathbf{k}}}(LU) \geq \frac{\gamma}{1+\gamma} \\ \hat{Q}_{n,h_{\mathbf{k}}}\left(\hat{F}_{n,h_{\mathbf{k}}}(LU)-\left(\frac{\gamma}{1+\gamma}-\hat{F}_{n,h_{\mathbf{k}}}(LU)\right)d\right) & \hat{F}_{n,h_{\mathbf{k}}}(LU) < \frac{\gamma}{1+\gamma}, \end{cases}$$

where LU is $LU_{\mathbf{k},k,\epsilon,\gamma,n}$, $\hat{F}_{n,h_{\mathbf{k}}}(x)$ is the empirical cumulative distribution function of the $h_{\mathbf{k}}$ kernel distribution, $\hat{Q}_{n,h_{\mathbf{k}}}$ is the quantile function of the $h_{\mathbf{k}}$ kernel distribution.

Similarly, the quantile mean can be defined as $qm_{d,k,\epsilon,\gamma,n,\mathrm{WL}} := \mathrm{QI}_{d,h_{\mathbf{k}}=x,\mathbf{k}=1,k,\epsilon,\gamma,n,LU=\mathrm{WL}}$. Moreover, in extreme right-skewed heavy-tailed distributions, if the calculated percentile exceeds $1-\epsilon$, it will be adjusted to $1-\epsilon$. In a left-skewed distribution, if the obtained percentile is smaller than $\gamma\epsilon$, it will also be adjusted to $\gamma\epsilon$. Without loss of generality, in the following discussion, only the case where $\hat{F}_n(\mathrm{WL}_{k,\epsilon,\gamma,n}) \geq \frac{\gamma}{1+\gamma}$ is considered. A widely used method for calculating the sample quantile function involves employing linear interpolation of modes corresponding to the order statistics of the uniform distribution on the interval [0, 1], i.e., $\hat{Q}_n(p) = X_{\lfloor h\rfloor}+(h-\lfloor h\rfloor)\left(X_{\lceil h\rceil}-X_{\lfloor h\rfloor}\right)$, $h = (n-1)p+1$. To minimize the finite sample bias, here, the inverse function of $\hat{Q}_n$ is deduced as $\hat{F}_n(x) := \frac{1}{n-1}\left(cf-1+\frac{x-X_{cf}}{X_{cf+1}-X_{cf}}\right)$, where $cf = \sum_{i=1}^{n}\mathbf{1}_{X_i\leq x}$, $\mathbf{1}_A$ is the indicator of event A. The quantile mean uses the location-scale invariant in a different way, as shown in the subsequent proof.

Theorem B.2. *$qm_{d=\frac{F(\mu)-F(WL_{k,\epsilon,\gamma})}{F(WL_{k,\epsilon,\gamma})-\frac{\gamma}{1+\gamma}},k,\epsilon,\gamma,\mathrm{WL}}$ is a consistent mean estimator for a location-scale distribution provided that the means are finite and $F(\mu)$, $F(WL_{k,\epsilon,\gamma})$ and $\frac{\gamma}{1+\gamma}$ are all within the range of $[\gamma\epsilon, 1-\epsilon]$, where μ and $WL_{k,\epsilon,\gamma}$ are location parameters from that location-scale distribution. If $WL = WHLM$, qm is also consistent for any γ-symmetric distributions.*

Proof. When $F(\mathrm{WL}_{k,\epsilon,\gamma}) \geq \frac{\gamma}{1+\gamma}$, the solution of $\left(F(\mathrm{WL}_{k,\epsilon,\gamma})-\frac{\gamma}{1+\gamma}\right)d + F(\mathrm{WL}_{k,\epsilon,\gamma}) = F(\mu)$ is $d = \frac{F(\mu)-F(\mathrm{WL}_{k,\epsilon,\gamma})}{F(\mathrm{WL}_{k,\epsilon,\gamma})-\frac{\gamma}{1+\gamma}}$. The d value for the case where $F(\mathrm{WL}_{k,\epsilon,\gamma,n}) < \frac{\gamma}{1+\gamma}$ is the same. The definitions of the location and scale parameters are such that they must satisfy $F(x;\lambda,\mu) = F(\frac{x-\mu}{\lambda};1,0)$, then $F(\mathrm{WL}(k,\epsilon,\gamma);\lambda,\mu) = F(\frac{\lambda\mathrm{WL}_0(k,\epsilon,\gamma)+\mu-\mu}{\lambda};1,0) = F(\mathrm{WL}_0(k,\epsilon,\gamma);1,0)$. It follows that the percentile of any weighted L-statistic is free of λ and μ for a location-scale distribution. Therefore d in qm is also invariably a constant. For the γ-symmetric case, $F(\mathrm{WHLM}_{k,\epsilon,\gamma}) = F(\mu) = F(Q(\frac{\gamma}{1+\gamma})) = \frac{\gamma}{1+\gamma}$ is valid for any γ-symmetric distribution with a finite second moment, as the same values correspond to same percentiles. Then, $qm_{d,k,\epsilon,\gamma,\mathrm{WHLM}} = F^{-1}\left(\left(F(\mathrm{WHLM}_{k,\epsilon,\gamma})-\frac{\gamma}{1+\gamma}\right)d+F(\mu)\right) = F^{-1}(0+F(\mu)) = \mu$. To avoid inconsistency due to post-adjustment, $F(\mu)$, $F(\mathrm{WL}_{k,\epsilon,\gamma})$ and $\frac{\gamma}{1+\gamma}$ must reside within the range of $[\gamma\epsilon, 1-\epsilon]$. All results are now proven. □

The cdf of the Pareto distribution is $F_{Par}(x) = 1-\left(\frac{x_m}{x}\right)^{\alpha}$. So, set the d value in qm with two arbitrary percentiles p_1 and p_2, $d_{Par,qm} = \frac{1-\left(\frac{x_m}{\frac{\alpha x_m}{\alpha-1}}\right)^{\alpha}-\left(1-\left(\frac{x_m}{x_m(1-p_1)^{-\frac{1}{\alpha}}}\right)^{\alpha}\right)}{\left(1-\left(\frac{x_m}{x_m(1-p_1)^{-\frac{1}{\alpha}}}\right)^{\alpha}\right)-\left(1-\left(\frac{x_m}{x_m(1-p_2)^{-\frac{1}{\alpha}}}\right)^{\alpha}\right)} = \frac{1-\left(\frac{\alpha-1}{\alpha}\right)^{\alpha}-p_1}{p_1-p_2}$. The d value in qm for the exponential distribution is always identical to $d_{Par,qm}$ as $\alpha \to \infty$, since $\lim_{\alpha\to\infty}\left(\frac{\alpha-1}{\alpha}\right)^{\alpha} = \frac{1}{e}$ and the cdf of the exponential distribution is $F_{exp}(x) = 1-e^{-\lambda^{-1}x}$, then $d_{exp,qm} = \frac{(1-e^{-1})-\left(1-e^{-\ln\left(\frac{1}{1-p_1}\right)}\right)}{\left(1-e^{-\ln\left(\frac{1}{1-p_1}\right)}\right)-\left(1-e^{-\ln\left(\frac{1}{1-p_2}\right)}\right)} = \frac{1-\frac{1}{e}-p_1}{p_1-p_2}$. So, for the Weibull, gamma, Pareto, lognormal and generalized Gaussian distribution, $qm_{d=\frac{F_{exp}(\mu)-F_{exp}(\mathrm{WHLM}_{k,\epsilon,\gamma})}{F_{exp}(\mathrm{WHLM}_{k,\epsilon,\gamma})-\frac{\gamma}{1+\gamma}},k,\epsilon,\gamma,\mathrm{WHLM}}$ is also consistent for at least one particular case, provided that μ and $\mathrm{WHLM}_{k,\epsilon,\gamma}$ are different location parameters from an exponential distribution and $F(\mu)$, $F(\mathrm{WHLM}_{k,\epsilon,\gamma})$ and $\frac{\gamma}{1+\gamma}$ are all within the range of $[\gamma\epsilon, 1-\epsilon]$. Also let $\mathrm{WHLM}_{k,\epsilon,\gamma} = \mathrm{BM}_{\nu=3,\epsilon=\frac{1}{24}}$ and $\mu = \lambda$, then $d = \frac{F_{exp}(\mu)-F_{exp}(\mathrm{BM}_{\nu=3,\epsilon=\frac{1}{24}})}{F_{exp}(\mathrm{BM}_{\nu=3,\epsilon=\frac{1}{24}})-\frac{1}{2}} =$

$$\frac{-e^{-1}+e^{-\left(1+\ln\left(\frac{26068394603446272\sqrt[6]{\frac{7}{247}}\sqrt[3]{11}}{391^{5/6}101898752449325\sqrt{5}}\right)\right)}}{\frac{1}{2}-e^{-\left(1+\ln\left(\frac{26068394603446272\sqrt[6]{\frac{7}{247}}\sqrt[3]{11}}{391^{5/6}101898752449325\sqrt{5}}\right)\right)}} = \frac{\frac{101898752449325\sqrt{5}\sqrt[6]{\frac{247}{7}}391^{5/6}}{26068394603446272\sqrt[3]{11}e}-\frac{1}{e}}{\frac{1}{2}-\frac{101898752449325\sqrt{5}\sqrt[6]{\frac{247}{7}}391^{5/6}}{26068394603446272\sqrt[3]{11}e}} \approx 0.088.$$

$F_{exp}(\mu)$, $F_{exp}(\mathrm{BM}_{\nu=3,\epsilon=\frac{1}{24}})$ and $\frac{1}{2}$ are all within the range of $[\frac{1}{24}, \frac{23}{24}]$. $qm_{d\approx0.088,\nu=3,\epsilon=\frac{1}{24},\mathrm{BM}}$ works better in the fat-tail scenarios (SI Dataset S1). Theorem B.1 and B.2 show that $rm_{d\approx0.103,\nu=3,\epsilon=\frac{1}{24},\mathrm{BM},m}$ and $qm_{d\approx0.088,\nu=3,\epsilon=\frac{1}{24},\mathrm{BM}}$ are both consistent mean estimators for any symmetric distribution and the exponential distribution with finite second moments. It's obvious that the asymptotic breakdown points of $rm_{d\approx0.103,\nu=3,\epsilon=\frac{1}{24},\mathrm{BM},m}$ and $qm_{d\approx0.088,\nu=3,\epsilon=\frac{1}{24},\mathrm{BM}}$ are both $\frac{1}{24}$. Therefore they are all invariant means.

To study the impact of the choice of WLs in rm and qm, it is constructive to recall that a weighted L-statistic is a combination of order statistics. While using a less-biased weighted L-statistic can generally enhance performance (SI Dataset S1), there is a greater risk of violation in the semiparametric framework. However, the mean-$\mathrm{WA}_{\epsilon,\gamma}$-$\gamma$-median inequality is robust to slight fluctuations of the QA function of the underlying distribution. Suppose for a right-skewed distribution, the QA function is generally decreasing with respect to ϵ in $[0, u]$, but increasing in $[u, \frac{1}{1+\gamma}]$, since all quantile averages with breakdown points from ϵ to $\frac{1}{1+\gamma}$ will be included in the computation of $\mathrm{WA}_{\epsilon,\gamma}$, as long as $\frac{1}{1+\gamma} - u \ll \frac{1}{1+\gamma} - \gamma\epsilon$, and other portions of the QA function satisfy the inequality constraints that define the νth γ-orderliness on which the $\mathrm{WA}_{\epsilon,\gamma}$ is based, if $0 \le \gamma \le 1$, the mean-$\mathrm{WA}_{\epsilon,\gamma}$-$\gamma$-median inequality still holds. This is due to the violation of νth γ-orderliness being bounded, when $0 \le \gamma \le 1$, as shown in RESM I and therefore cannot be extreme for unimodal distributions with finite second moments. For instance, the SQA function of the Weibull distribution is non-monotonic with respect to ϵ when the shape parameter $\alpha > \frac{1}{1-\ln(2)} \approx 3.259$ as shown in the SI Text of RESM I, the violation of the second and third orderliness starts near this parameter as well, yet the mean-$\mathrm{BM}_{\nu=3,\epsilon=\frac{1}{24}}$-median inequality retains valid when $\alpha \le 3.387$. Another key factor in determining the risk of violation of orderliness is the skewness of the distribution. In RESM I, it was demonstrated that in a family of distributions differing by a skewness-increasing transformation in van Zwet's sense, the violation of orderliness, if it happens, only occurs as the distribution nears symmetry (12). When $\gamma = 1$, the over-corrections in rm and qm are dependent on the SWA_{ϵ}-median difference, which can be a reasonable measure of skewness after standardization (11, 13), implying that the over-correction is often tiny with moderate d. This qualitative analysis suggests the general reliability of rm and qm based on the mean-$\mathrm{WA}_{\epsilon,\gamma}$-$\gamma$-median inequality, especially for unimodal distributions with finite second moments when $0 \le \gamma \le 1$. Extending this rationale to other weighted L-statistics is possible, since the γ-U-orderliness can also be bounded with certain assumptions, as discussed previously.

Another crucial property of the central moment kernel distribution, location invariant, is introduced in the next theorem. The proof is provided in the SI Text.

Theorem B.3. $\psi_{\mathbf{k}}(x_1 = \lambda x_1 + \mu, \cdots, x_{\mathbf{k}} = \lambda x_{\mathbf{k}} + \mu) = \lambda^{\mathbf{k}}\psi_{\mathbf{k}}(x_1, \cdots, x_{\mathbf{k}})$.

A direct result of Theorem B.3 is that, WHL**k**m after standardization is invariant to location and scale. So, the weighted H-L standardized **k**th moment is defined to be

$$\mathrm{WHL}s\mathbf{k}m_{\epsilon=\min(\epsilon_1,\epsilon_2),k_1,k_2,\gamma_1,\gamma_2,n} \coloneqq \frac{\mathrm{WHL}\mathbf{k}m_{k_1,\epsilon_1,\gamma_1,n}}{(\mathrm{WHL}var_{k_2,\epsilon_2,\gamma_2,n})^{\mathbf{k}/2}}.$$

Consider two continuous distributions belonging to the same location–scale family, according to Theorem B.3, their corresponding **k**th central moment kernel distributions only differ in scaling. Define the recombined **k**th central moment as $r\mathbf{k}m_{d,k_1,k_2,\epsilon=\min(\epsilon_1,\epsilon_2),\gamma_1,\gamma_2,n,\mathrm{WHL}\mathbf{k}m_1,\mathrm{WHL}\mathbf{k}m_2} \coloneqq \mathrm{RI}_{d,h_{\mathbf{k}}=\psi_{\mathbf{k}},\mathbf{k}_1=\mathbf{k},\mathbf{k}_2=\mathbf{k},k_1,k_2,\epsilon_1,\epsilon_2,\gamma_1,\gamma_2,n,LU_1=\mathrm{WHL}\mathbf{k}m_1,LU_2=\mathrm{WHL}\mathbf{k}m_2}$. Then, assuming finite **k**th central moment and applying the same logic as in Theorem B.1, $r\mathbf{k}m_{d=\frac{\mu_{\mathbf{k}}-\mathrm{WHL}\mathbf{k}m_{1k_1,\epsilon_1,\gamma_1}}{\mathrm{WHL}\mathbf{k}m_{1k_1,\epsilon_1,\gamma_1}-\mathrm{WHL}\mathbf{k}m_{2k_2,\epsilon_2,\gamma_2}},k_1,k_2,\epsilon=\min(\epsilon_1,\epsilon_2),\gamma_1,\gamma_2,\mathrm{WHL}\mathbf{k}m_1,\mathrm{WHL}\mathbf{k}m}$ is a consistent **k**th central moment estimator for a location-scale distribution, where $\mu_{\mathbf{k}}$, $\mathrm{WHL}\mathbf{k}m_{1k_1,\epsilon_1,\gamma_1}$, and $\mathrm{WHL}\mathbf{k}m_{2k_2,\epsilon_2,\gamma_2}$ are different **k**th central moment parameters from that location-scale distribution. Similarly, the quantile will not change after scaling. The quantile **k**th central moment is thus defined as

$$q\mathbf{k}m_{d,k,\epsilon,\gamma,n,\mathrm{WHL}\mathbf{k}m} \coloneqq \mathrm{QI}_{d,h_{\mathbf{k}}=\psi_{\mathbf{k}},\mathbf{k}=\mathbf{k},k,\epsilon,\gamma,n,LU=\mathrm{WHL}\mathbf{k}m}.$$

$q\mathbf{k}m_{d=\frac{F_{\psi_{\mathbf{k}}}(\mu_{\mathbf{k}})-F_{\psi_{\mathbf{k}}}(\mathrm{WHL}\mathbf{k}m_{k,\epsilon,\gamma})}{F_{\psi_{\mathbf{k}}}(\mathrm{WHL}\mathbf{k}m_{k,\epsilon,\gamma})-\frac{\gamma}{1+\gamma}},k,\epsilon,\gamma,\mathrm{WHL}\mathbf{k}m}$ is also a consistent **k**th central moment estimator for a location-scale distribution provided that the **k**th central moment is finite and $F_{\psi_{\mathbf{k}}}(\mu_{\mathbf{k}})$, $F_{\psi_{\mathbf{k}}}(\mathrm{WHL}\mathbf{k}m_{k,\epsilon,\gamma})$ and $\frac{\gamma}{1+\gamma}$ are all within the range of $[\gamma\epsilon, 1-\epsilon]$, where $\mu_{\mathbf{k}}$ and $\mathrm{WHL}\mathbf{k}m_{k,\epsilon,\gamma}$ are different **k**th central moment parameters from that location-scale distribution.

So, the quantile standardized **k**th moment is defined to be

$$qs\mathbf{k}m_{\epsilon=\min(\epsilon_1,\epsilon_2),k_1,k_2,\gamma_1,\gamma_2,n,\mathrm{WHL}\mathbf{k}m,\mathrm{WHL}var} \coloneqq \frac{q\mathbf{k}m_{d,k_1,\epsilon_1,\gamma_1,n,\mathrm{WHL}\mathbf{k}m}}{(qvar_{d,k_2,\epsilon_2,\gamma_2,n,\mathrm{WHL}var})^{\mathbf{k}/2}}.$$

The recombined standardized **k**th moment ($rs\mathbf{k}m_{\epsilon=\min(\epsilon_1,\epsilon_2),k_1,k_2,\gamma_1,\gamma_2,n,\mathrm{WHL}\mathbf{k}m_1,\mathrm{WHL}\mathbf{k}m_2,\mathrm{WHL}var_1,\mathrm{WHL}var_2}$) is defined similarly and not repeated here. From the better performance of the quantile mean in heavy-tailed distributions, the quantile **k**th central moments are generally better than recombined **k**th central moments regarding asymptotic bias.

C. Congruent Distribution. In the realm of nonparametric statistics, the relative differences, or orders, of robust estimators are of primary importance. A key implication of this principle is that when there is a shift in the parameters of the underlying distribution, all nonparametric estimates should asymptotically change in the same direction, if they are estimating the same attribute of the distribution. If, on the other hand, the mean suggests an increase in the location of the distribution while the median indicates a decrease, a contradiction arises. It is worth noting that such contradiction is not possible for any LL-statistics in a location-scale distribution, as explained in the previous article on semiparametric robust mean. However, it is possible to construct counterexamples to the aforementioned implication in a shape-scale distribution. In the case of the Weibull distribution, its quantile function is $Q_{Wei}(p) = \lambda(-\ln(1-p))^{1/\alpha}$, where $0 \le p \le 1$, $\alpha > 0$, $\lambda > 0$, λ is a scale parameter, α is a shape parameter, ln is the natural logarithm function. Then,

$m = \lambda \sqrt[\alpha]{\ln(2)}$, $\mu = \lambda\Gamma\left(1+\frac{1}{\alpha}\right)$, where Γ is the gamma function. When $\alpha = 1$, $m = \lambda\ln(2) \approx 0.693\lambda$, $\mu = \lambda$, when $\alpha = \frac{1}{2}$, $m = \lambda\ln^2(2) \approx 0.480\lambda$, $\mu = 2\lambda$, the mean increases as α changes from 1 to $\frac{1}{2}$, but the median decreases. Previously, the fundamental role of quantile average and its relation to nearly all common nonparametric robust location estimates were demonstrated by using the method of classifying distributions through the signs of derivatives. To avoid such scenarios, this method can also be used. Let the quantile average function of a parametric distribution be denoted as $\text{QA}\left(\epsilon, \gamma, \alpha_1, \cdots, \alpha_i, \cdots, \alpha_k\right)$, where α_i represent the parameters of the distribution, then, a distribution is γ-congruent if and only if the sign of $\frac{\partial \text{QA}}{\partial \alpha_i}$ remains the same for all $0 \le \epsilon \le \frac{1}{1+\gamma}$. If $\frac{\partial \text{QA}}{\partial \alpha_i}$ is equal to zero or undefined, it can be considered both positive and negative, and thus does not impact the analysis. A distribution is completely γ-congruent if and only if it is γ-congruent and all its central moment kernel distributions are also γ-congruent. Setting $\gamma = 1$ constitutes the definitions of congruence and complete congruence. Replacing the QA with γmHLM gives the definition of γ-U-congruence. Chebyshev's inequality implies that, for any probability distributions with finite second moments, as the parameters change, even if some LL-statistics change in a direction different from that of the population mean, the magnitude of the changes in the LL-statistics remains bounded compared to the changes in the population mean. Furthermore, distributions with infinite moments can be γ-congruent, since the definition is based on the quantile average, not the population mean.

The following theorems show the conditions that a distribution is congruent or γ-congruent.

Theorem C.1. *A γ-symmetric distribution is always γ-congruent and γ-U-congruent.*

Proof. As shown in RESM I, Theorem .2 and Theorem .18, for any γ-symmetric distribution, all quantile averages and all γmHLMs conincide. The conclusion follows immediately. ☐

Theorem C.2. *A positive definite location-scale distribution is always γ-congruent.*

Proof. As shown in RESM I, Theorem .2, for a location-scale distribution, any quantile average can be expressed as $\lambda \text{QA}_0(\epsilon, \gamma) + \mu$. Therefore, the derivatives with respect to the parameters λ or μ are always positive. By application of the definition, the desired outcome is obtained. ☐

Theorem C.3. *The second central moment kernal distribution derived from a continuous location-scale unimodal distribution is always γ-congruent.*

Proof. Theorem B.3 shows that the central moment kernel distribution generated from a location-scale distribution is also a location-scale distribution. Theorem A.1 shows that it is positively definite. Implementing Theorem C.2 yields the desired result. ☐

For the Pareto distribution, $\frac{\partial Q}{\partial \alpha} = \frac{x_m(1-p)^{-1/\alpha}\ln(1-p)}{\alpha^2}$. Since $\ln(1-p) < 0$ for all $0 < p < 1$, $(1-p)^{-1/\alpha} > 0$ for all $0 < p < 1$ and $\alpha > 0$, so $\frac{\partial Q}{\partial \alpha} < 0$, and therefore $\frac{\partial \text{QA}}{\partial \alpha} < 0$, the Pareto distribution is γ-congruent. It is also γ-U-congruent, since γmHLM can also express as a function of $Q(p)$. For the lognormal distribution, $\frac{\partial \text{QA}}{\partial \sigma} = \frac{1}{2}\left(\sqrt{2}\text{erfc}^{-1}(2\gamma\epsilon)\left(-e^{\frac{\sqrt{2}\mu - 2\sigma\text{erfc}^{-1}(2\gamma\epsilon)}{\sqrt{2}}}\right) + \left(-\sqrt{2}\right)\text{erfc}^{-1}(2(1-\epsilon))e^{\frac{\sqrt{2}\mu - 2\sigma\text{erfc}^{-1}(2(1-\epsilon))}{\sqrt{2}}}\right)$. Since the inverse complementary error function is positive when the input is smaller than 1, and negative when the input is larger than 1, and symmetry around 1, if $0 \le \gamma \le 1$, $\text{erfc}^{-1}(2\gamma\epsilon) \ge -\text{erfc}^{-1}(2-2\epsilon)$, $e^{\mu-\sqrt{2}\sigma\text{erfc}^{-1}(2-2\epsilon)} > e^{\mu-\sqrt{2}\sigma\text{erfc}^{-1}(2\gamma\epsilon)}$. Therefore, if $0 \le \gamma \le 1$, $\frac{\partial \text{QA}}{\partial \sigma} > 0$, the lognormal distribution is γ-congruent. Theorem C.1 implies that the generalized Gaussian distribution is congruent and U-congruent. For the Weibull distribution, when α changes from 1 to $\frac{1}{2}$, the average probability density on the left side of the median increases, since $\frac{\frac{1}{2}}{\lambda\ln(2)} < \frac{\frac{1}{2}}{\lambda\ln^2(2)}$, but the mean increases, indicating that the distribution is more heavy-tailed, the probability density of large values will also increase. So, the reason for non-congruence of the Weibull distribution lies in the simultaneous increase of probability densities on two opposite sides as the shape parameter changes: one approaching the bound zero and the other approaching infinity. Note that the gamma distribution does not have this issue, Numerical results indicate that it is likely to be congruent.

Although some parametric distributions are not congruent, Theorem C.2 establishes that γ-congruence always holds for a positive definite location-scale family distribution and thus for the second central moment kernel distribution generated from a location-scale unimodal distribution as shown in Theorem C.3. Theorem A.2 demonstrates that all central moment kernel distributions are unimodal-like with mode and median close to zero, as long as they are generated from unimodal distributions. Assuming finite moments and constant $Q(0) - Q(1)$, increasing the mean of a distribution will result in a generally more heavy-tailed distribution, i.e., the probability density of the values close to $Q(1)$ increases, since the total probability density is 1. In the case of the $\mathbf{k}$th central moment kernel distribution, $\mathbf{k} > 2$, while the total probability density on either side of zero remains generally constant as the median is generally close to zero and much less impacted by increasing the mean, the probability density of the values close to zero decreases as the mean increases. This transformation will increase nearly all symmetric weighted averages, in the general sense. Therefore, except for the median, which is assumed to be zero, nearly all symmetric weighted averages for all central moment kernel distributions derived from unimodal distributions should change in the same direction when the parameters change.

D. A Shape-Scale Distribution as the Consistent Distribution. In Subsection B, the parametric robust estimation is limited to a location-scale distribution, with the location parameter often being omitted for simplicity. For improved fit to observed skewness or kurtosis, shape-scale distributions with shape parameter (α) and scale parameter (λ) are commonly utilized. Weibull, gamma, Pareto, lognormal, and generalized Gaussian distributions (when μ is a constant) are all shape-scale unimodal distributions. Furthermore, if either the shape parameter α or the skewness or kurtosis is constant, the shape-scale distribution is reduced to a location-scale distribution. Let $D(|skewness|, kurtosis, \mathbf{k}, etype, dtype, n) = d_{i\mathbf{k}m}$ denote the function to specify d values, where the first input is the absolute value of the skewness, the second input is the kurtosis, the third is the order of the central moment (if $\mathbf{k}=1$, the mean), the fourth is the type of estimator, the fifth is the type of consistent distribution, and the sixth input is the sample size. For simplicity, the last three inputs will be omitted in the following discussion. Hold in awareness that since skewness and kurtosis are interrelated, specifying d values for a shape-scale distribution only requires either skewness or kurtosis, while the other may be also omitted. Since many common shape-scale distributions are always right-skewed (if not, only the right-skewed or left-skewed part is used for calibration, while the other part is omitted), the absolute value of the skewness should be the same as the skewness of these distributions. This setting also handles the left-skew scenario well.

For recombined moments up to the fourth ordinal, the object of using a shape-scale distribution as the consistent distribution is to find solutions for the system of equations
$$\begin{cases} rm\left(\mathrm{WL}, \gamma m, D(|rskew|, rkurt, 1)\right) = \mu \\ rvar\left(\mathrm{WHL}var, \gamma mvar, D(|rskew|, rkurt, 2)\right) = \mu_2 \\ rtm\left(\mathrm{WHL}tm, \gamma mtm, D(|rskew|, rkurt, 3)\right) = \mu_3 \\ rfm\left(\mathrm{WHL}fm, \gamma mfm, D(|rskew|, rkurt, 4\right) = \mu_4 \\ rskew = \frac{\mu_3}{\mu_2^{\frac{3}{2}}} \\ rkurt = \frac{\mu_4}{\mu_2^2} \end{cases},$$
where μ_2, μ_3 and μ_4 are the population second, third and fourth central moments. $|rskew|$ and $rkurt$ should be the invariant points of the functions $\varsigma(|rskew|) = \left|\frac{rtm(\mathrm{WHL}tm, \gamma mtm, D(|rskew|, 3))}{rvar(\mathrm{WHL}var, \gamma mvar, D(|rskew|, 2))^{\frac{3}{2}}}\right|$ and $\varkappa(rkurt) = \frac{rfm(\mathrm{WHL}fm, \gamma mfm, D(rkurt, 4))}{rvar(\mathrm{WHL}var, \gamma mvar, D(rkurt, 2))^2}$. Clearly, this is an overdetermined nonlinear system of equations, given that the skewness and kurtosis are interrelated for a shape-scale distribution. Since an overdetermined system constructed with random coefficients is almost always inconsistent, it is natural to optimize them separately using the fixed-point iteration (see Algorithm 1, only $rkurt$ is provided, others are the same).

Algorithm 1 $rkurt$ for a shape-scale distribution

Input: D; WHLvar; WHLfm; $\gamma mvar$; γmfm; $maxit$; δ

Output: $rkurt_{i-1}$

$i = 0$

$rkurt_i \leftarrow \varkappa(kurtosis_{max})$ ▷ Using the maximum kurtosis available in D as an initial guess.

repeat

  $i = i + 1$

  $rkurt_{i-1} \leftarrow rkurt_i$

  $rkurt_i \leftarrow \varkappa(rkurt_{i-1})$

until $i > maxit$ or $|rkurt_i - rkurt_{i-1}| < \delta$ ▷ $maxit$ is the maximum number of iterations, δ is a small positive number.

The following theorem shows the validity of Algorithm 1.

Theorem D.1. *Assuming $\gamma = 1$ and $m\mathbf{k}ms$, where $2 \leq \mathbf{k} \leq 4$, are all equal to zero, $|rskew|$ and $rkurt$, defined as the largest attracting fixed points of the functions $\varsigma(|rskew|)$ and $\varkappa(rkurt)$, are consistent estimators of $\tilde{\mu}_3$ and $\tilde{\mu}_4$ for a shape-scale distribution whose* **k***th central moment kernel distributions are γ-U-congruent, as long as they are within the domain of D, where $\tilde{\mu}_3$ and $\tilde{\mu}_4$ are the population skewness and kurtosis, respectively.*

Proof. Without loss of generality, only $rkurt$ is considered, while the logic for $|rskew|$ is the same. Additionally, the second central moments of the underlying sample distribution and consistent distribution are assumed to be 1, with other cases simply multiplying a constant factor according to Theorem B.3. From the definition of D, $\frac{\varkappa(rkurt_D)}{rkurt_D} = \frac{\frac{fm_D - \mathrm{SWHL}fm_D}{\mathrm{SWHL}fm_D - mfm_D}(\mathrm{SWHL}fm - mfm) + \mathrm{SWHL}fm}{rkurt_D\left(\frac{var_D - \mathrm{SWHL}var_D}{\mathrm{SWHL}var_D - mvar_D}(\mathrm{SWHL}var - mvar) + \mathrm{SWHL}var\right)^2}$, where the subscript D indicates that the estimates are from the central moment kernel distributions generated from the consistent distribution, while other estimates are from the underlying distribution of the sample.

Then, assuming the $m\mathbf{k}ms$ are all equal to zero and $var_D = 1$, $\frac{\varkappa(rkurt_D)}{rkurt_D} = \frac{\frac{fm_D - \mathrm{SWHL}fm_D}{\mathrm{SWHL}fm_D}(\mathrm{SWHL}fm) + \mathrm{SWHL}fm}{rkurt_D\left(\frac{\mathrm{SWHL}var}{\mathrm{SWHL}var_D}\right)^2} = \frac{\left(\frac{fm_D - \mathrm{SWHL}fm_D}{\mathrm{SWHL}fm_D} + 1\right)(\mathrm{SWHL}fm)}{fm_D\left(\frac{\mathrm{SWHL}var}{\mathrm{SWHL}var_D}\right)^2} = \frac{\mathrm{SWHL}fm\mathrm{SWHL}var_D^2}{\mathrm{SWHL}fm_D\mathrm{SWHL}var^2} = \frac{\frac{\mathrm{SWHL}fm}{\mathrm{SWHL}var^2}}{\frac{\mathrm{SWHL}fm_D}{\mathrm{SWHL}var_D{}^2}} = \frac{\mathrm{SWHL}kurt}{\mathrm{SWHL}kurt_D}$. Since SWHL$fm_D$ are from the same fourth central moment kernel distribution as $fm_D = rkurt_D var_D{}^2$, according to the definition of γ-U-congruence, an increase in fm_D will also result in an increase in SWHLfm_D. Combining with Theorem B.3, SWHL$kurt$ is a measure of kurtosis that is invariant to location and scale, so $\lim_{rkurt_D \to \infty} \frac{\varkappa(rkurt_D)}{rkurt_D} < 1$. As a result, if there is at least one fixed point, let the largest one be fix_{max}, then it is attracting since $|\frac{\partial(\varkappa(rkurt_D))}{\partial(rkurt_D)}| < 1$ for all $rkurt_D \in [fix_{max}, kurtosis_{max}]$, where $kurtosis_{max}$ is the maximum kurtosis available in D.

□

As a result of Theorem D.1, assuming continuity, $m\mathbf{k}ms$ are all equal to zero, and γ-U-congruence of the central moment kernel distributions, Algorithm 1 converges surely provided that a fixed point exists within the domain of D. At this stage, D can only be approximated through a Monte Carlo study. The continuity of D can be ensured by using linear interpolation. One common encountered problem is that the domain of D depends on both the consistent distribution and the Monte Carlo study, so the iteration may halt at the boundary if the fixed point is not within the domain. However, by setting a proper maximum number of iterations, the algorithm can return the optimal boundary value. For quantile moments, the logic is similar, if the percentiles do not exceed the breakdown point. If this is the case, consistent estimation is impossible, and the algorithm will stop due to the maximum number of iterations. The fixed point iteration is, in principle, similar to the iterative reweighing in Huber M-estimator, but an advantage of this algorithm is that the optimization is solely related to the inputs in Algorithm 1 and is independent of the sample size. Since $|rskew|$ and $rkurt$ can specify d_{rm} and d_{rvar} after optimization, this algorithm enables the robust estimations of all four moments to reach a near-consistent level for common unimodal distributions (Table 1, SI Dataset S1), just using the Weibull distribution as the consistent distribution.

E. Variance. As one of the fundamental theorems in statistics, the Central Limit Theorem declares that the standard deviation of the limiting form of the sampling distribution of the sample mean is $\frac{\sigma}{\sqrt{n}}$. The principle, asymptotic normality, was later applied to the sampling distributions of robust location estimators. Bickel and Lehmann, also in the landmark series (18, 29), argued that meaningful comparisons of the efficiencies of various kinds of location estimators can be accomplished by studying their standardized variances, asymptotic variances, and efficiency bounds. Standardized variance, $\frac{\mathrm{Var}(\hat{\theta})}{\theta^2}$, allows the use of simulation studies or empirical data to compare the variances of estimators of distinct parameters. However, a limitation of this approach is the inverse square dependence of the standardized variance on θ. If $\mathrm{Var}\left(\hat{\theta}_1\right) = \mathrm{Var}\left(\hat{\theta}_2\right)$, but θ_1 is close to zero and θ_2 is relatively large, their standardized variances will still differ dramatically. Here, the scaled standard error (SSE) is proposed as a method for estimating the variances of estimators measuring the same attribute, offering a standard error more comparable to that of the sample mean and much less influenced by the magnitude of θ.

Definition E.1 (Scaled standard error). Let $\mathcal{M}_{s_i s_j} \in \mathbb{R}^{i \times j}$ denote the sample-by-statistics matrix, i.e., the first column corresponds to $\widehat{\theta_U}$, which is the mean or a U-central moment measuring the same attribute of the distribution as the other columns, the second to the jth column correspond to $j-1$ statistics required to scale, $\widehat{\theta_{r_1}}$, $\widehat{\theta_{r_2}}$, ..., $\widehat{\theta_{r_{j-1}}}$. Then, the scaling factor $\mathcal{S} = \left[1, \frac{\bar{\theta_{r_1}}}{\bar{\theta_m}}, \frac{\bar{\theta_{r_2}}}{\bar{\theta_m}}, \ldots, \frac{\bar{\theta_{r_{j-1}}}}{\bar{\theta_m}}\right]^T$ is a $j \times 1$ matrix, which $\bar{\theta}$ is the mean of the column of $\mathcal{M}_{s_i s_j}$. The normalized matrix is $\mathcal{M}^N_{s_i s_j} = \mathcal{M}_{s_i s_j}\mathcal{S}$. The SSEs are the unbiased standard deviations of the corresponding columns of $\mathcal{M}^N_{s_i s_j}$.

The U-central moment (the central moment estimated by using U-statistics) is essentially the mean of the central moment kernel distribution, so its standard error should be generally close to $\frac{\sigma_{km}}{\sqrt{n}}$, although not exactly since the kernel distribution is not i.i.d., where σ_{km} is the asymptotic standard deviation of the central moment kernel distribution. If the statistics of interest coincide asymptotically, then the standard errors should still be used, e.g, for symmetric location estimators and odd ordinal central moments for the symmetric distributions, since the scaled standard error will be too sensitive to small changes when they are zero.

The SSEs of all robust estimators proposed here are often, although many exceptions exist, between those of the sample median and those of the sample mean or median central moments and U-central moments (SI Dataset S1). This is because similar monotonic relations between breakdown point and variance are also very common, e.g., Bickel and Lehmann (18) proved that a lower bound for the efficiency of TM_ϵ to sample mean is $(1-2\epsilon)^2$ and this monotonic bound holds true for any distribution. However, the direction of monotonicity differs for distributions with different kurtosis. Lehmann and Scheffé (1950, 1955) (30, 31) in their two early papers provided a way to construct a uniformly minimum-variance unbiased estimator (UMVUE). From that, the sample mean and unbiased sample second moment can be proven as the UMVUEs for the population mean and population second moment for the Gaussian distribution. While their performance for sub-Gaussian distributions is generally satisfied, they perform poorly when the distribution has a heavy tail and completely fail for distributions with infinite second moments. Therefore, for sub-Gaussian distributions, the variance of a robust location estimator is generally monotonic increasing as its robustness increases, but for heavy-tailed distributions, the relation is reversed. As a result, unlike bias, the variance-optimal choice can be very different for distributions with different kurtosis.

Lai, Robbins, and Yu (1983) proposed an estimator that adaptively chooses the mean or median in a symmetric distribution and showed that the choice is typically as good as the better of the sample mean and median regarding variance (32). Another approach can be dated back to Laplace (1812) (33) is using $w\bar{x} + (1-w)m_n$ as a location estimator and w is deduced to achieve optimal variance. In this study, for $rkurt$, there are 364 combinations based on 14 SWfms and 26 SWvars (SI Text). Each combination has a root mean square error (RMSE) for a single-parameter distribution, which can be inferred through a Monte Carlo study. For $qkurt$, there are another 364 combinations, but if the percentiles of quantile moments exceed the breakdown point, that combination is excluded. Then, the combination with the smallest RMSE is chosen. Similar to Subsection D, let $I(kurtosis, dtype, n) = ikurt_{\mathrm{SW}fm,\mathrm{SW}var}$ denote these relations (the breakdown points of the SWLs in SWkm were adjusted to ensure the overall breakdown points were $\frac{1}{24}$, as detailed in the SI Text). Since $\lim_{ikurt\to\infty} \frac{I(ikurt)}{ikurt} < 1$, the same fix point iteration algorithm can be used to choose the variance-optimum combination. The only difference is that unlike D, I is defined to be discontinuous but linear interpolation can also ensure continuity. The procedure for $iskew$ is the same. The RMSEs of rkm and qkm can also be estimated by a Monte Carlo study and the estimator with the smallest RMSE of each ordinal is named as ikm. $iskew$ and $ikurt$ are then used to determine ikm. This approach yields results that are often nearly optimal (SI Dateset S1)

Due to combinatorial explosion, the bootstrap (34), introduced by Efron in 1979, is indispensable for computing invariant central moments in practice. In 1981, Bickel and Freedman (35) showed that the bootstrap is asymptotically valid to approximate the original distribution in a wide range of situations, including U-statistics. The limit laws of bootstrapped trimmed U-statistics were proven by Helmers, Janssen, and Veraverbeke (1990) (36). In the previous article, the advantages of quasi-bootstrap were discussed (37–39). By using quasi-sampling, the impact of the number of repetitions of the bootstrap, or bootstrap size, on variance is very small (SI Dataset S1). An estimator based on the quasi-bootstrap approach can be seen as a complex deterministic estimator that is not only computationally efficient but also statistical efficient. The only drawback of quasi-bootstrap compared to non-bootstrap is that a small bootstrap size can produce additional finite sample bias (SI Text). The d values should be re-calibrated. In general, the variances of invariant central moments are much smaller than those of corresponding unbiased sample central moments (deduced by Cramér (40)), except that of the corresponding second central moment (Table 1).

F. Robustness. The measure of robustness to gross errors used in this series is the breakdown point proposed by Hampel (41) in 1968. In RESM I, it has shown that the median of means (MoM) is asymptotically equivalent to the median Hodge-Lehmann mean. Therefore it is also biased for any

Table 1. Evaluation of invariant moments for five common unimodal distributions in comparison with current popular methods

Errors	HM	$\bar{x}$	PE$_{\mu}$	im_v	Tsd^2	var	PE$_{\mu_2}$	$ivar_v$	tm	PE$_{\mu_3}$	itm_v	fm	PE$_{\mu_4}$	ifm_v
WASAB	0.102	0.000	0.048	0.002	0.234	0.000	0.072	0.047	0.000	0.099	0.013	0.000	0.115	0.109
WRMSE	0.106	0.016	0.064	0.016	0.233	0.019	0.097	0.052	0.023	0.124	0.021	0.029	0.151	0.118
WASB$_{n=4096}$	0.102	0.000	0.049	0.002	0.233	0.001	0.074	0.037	0.001	0.104	0.011	0.001	0.125	0.100
WSE $\vee$ WSSE	0.016	0.016	0.026	0.016	0.016	0.019	0.039	0.025	0.022	0.063	0.015	0.027	0.032	0.025

This table presents the use of the Weibull distribution as the consistent distribution plus optimization (ikm_v is invariant kth moment, variance-optimized) for five common unimodal distributions: Weibull, gamma, Pareto, lognormal and generalized Gaussian distributions. Unbiased sample moments, Huber M-estimator, and percentile estimator (PE) for the Weibull distribution (7) were used as comparisons. The Gaussian distribution was excluded for PE, since the logarithmic function does not produce results for negative inputs. The breakdown points of invariant moments are all $\frac{1}{24}$. The table includes the average standardized asymptotic bias (ASAB, as $n \to \infty$), root mean square error (RMSE, at $n = 4096$), average standardized bias (ASB, at $n = 4096$) and variance (SE $\vee$ SSE, at $n = 4096$) of these estimators, all reported in the units of the standard deviations of the distribution or corresponding kernel distributions. The notation bs indicates the quasi-bootstrap central moments. W means that the results were weighted by the number of Google Scholar search results (including synonyms). The calibrations of d values and the computations of ASAB, ASB, and SSE were described in Subsection E, F and SI Methods. Detailed results and related codes are available in SI Dataset S1.

asymmetric distribution. However, the concentration bound of MoM depends on $\sqrt{\frac{1}{n}}$ (42), it is quite natural to deduce that it is a consistent robust estimator. The concept, sample-dependent breakdown point, is defined to avoid ambiguity.

Definition F.1 (Sample-dependent breakdown point). The breakdown point of an estimator $\hat{\theta}$ is called sample-dependent if and only if the upper and lower asymptotic breakdown points, which are the upper and lower breakdown points when $n \to \infty$, are zero and the empirical influence function of $\hat{\theta}$ is bounded. For a full formal definition of the empirical influence function, the reader is referred to Devlin, Gnanadesikan and Kettenring (1975)'s paper (43).

Bear in mind that it differs from the "infinitesimal robustness" defined by Hampel, which is related to whether the asymptotic influence function is bounded (44–46). The proof of the consistency of MoM assumes that it is an estimator with a sample-dependent breakdown point since its breakdown point is $\frac{b}{2n}$, where b is the number of blocks, then $\lim_{n\to\infty}\left(\frac{b}{2n}\right) = 0$, if b is a constant and any changes in any one of the points of the sample cannot break down this estimator.

For the robust estimations of central moments or other LU-statistics, the asymptotic upper breakdown points are suggested by the following theorem, which extends the method in Donoho and Huber (1983)'s proof of the breakdown point of the Hodges-Lehmann estimator (47). The proof is given in the SI Text.

Theorem F.1. *Given a U-statistic associated with a symmetric kernel of degree* $\mathbf{k}$. *Then, assuming that as* $n \to \infty$, $\mathbf{k}$ *is a constant, the upper breakdown point of the LU-statistic is* $1 - (1 - \epsilon_{\mathbf{0}})^{\frac{1}{\mathbf{k}}}$, *where* $\epsilon_{\mathbf{0}}$ *is the upper breakdown point of the corresponding LL-statistic.*

Remark. If $\mathbf{k} = 1$, $1 - (1 - \epsilon_{\mathbf{0}})^{\frac{1}{\mathbf{k}}} = \epsilon_{\mathbf{0}}$, so this formula also holds for the LL-statistic itself. Here, to ensure the breakdown points of all four moments are the same, $\frac{1}{24}$, since $\epsilon_{\mathbf{0}} = 1 - (1 - \epsilon)^{\mathbf{k}}$, the breakdown points of all LU-statistics for the second, third, and fourth central moment estimations are adjusted as $\epsilon_{\mathbf{0}} = \frac{47}{576}, \frac{1657}{13824}, \frac{51935}{331776}$, respectively.

Every statistic is based on certain assumptions. For instance, the sample mean assumes that the second moment of the underlying distribution is finite. If this assumption is violated, the variance of the sample mean becomes infinitely large, even if the population mean is finite. As a result, the sample mean not only has zero robustness to gross errors, but also has zero robustness to departures. To meaningfully compare the performance of estimators under departures from assumptions, it is necessary to impose constraints on these departures. Bound analysis (1) is the first approach to study the robustness to departures, i.e., although all estimators can be biased under departures from the corresponding assumptions, but their standardized maximum deviations can differ substantially (42, 48–51). In RESM I, it is shown that another way to qualitatively compare the estimators' robustness to departures from the γ-symmetry assumption is constructing and comparing corresponding semiparametric models. While such comparison is limited to a semiparametric model and is not universal, it is still valid for a wide range of parametric distributions. Bound analysis is a more universal approach since they can be deduced by just assuming regularity conditions (42, 48, 49, 51). However, bounds are often hard to deduce for complex estimators. Also, sometimes there are discrepancies between maximum bias and average bias. Since the estimators proposed here are all consistent under certain assumptions, measuring their biases is also a convenient way of measuring the robustness to departures. Average standardized asymptotic bias is thus defined as follows.

Definition F.2 (Average standardized asymptotic bias). For a single-parameter distribution, the average standardized asymptotic bias (ASAB) is given by $\frac{|\hat{\theta}-\theta|}{\sigma}$, where $\hat{\theta}$ represents the estimation of θ, and σ denotes the standard deviation of the kernel distribution associated with the LU-statistic. If the estimator $\hat{\theta}$ is not classified as an RI-statistic, QI-statistic, or LU-statistic, the corresponding U-statistic, which measures the same attribute of the distribution, is utilized to determine the value of σ. For a two-parameter distribution, the first step is setting the lower bound of the kurtosis range of interest $\tilde{\mu}_{4_l}$, the spacing δ, and the bin count C. Then, the average standardized asymptotic bias is defined as

$$\mathrm{ASAB}_{\hat{\theta}} \coloneqq \frac{1}{C} \sum_{\substack{\delta+\tilde{\mu}_{4_l} \leq \tilde{\mu}_4 \leq C\delta+\tilde{\mu}_{4_l} \\ \tilde{\mu}_4 \text{ is a multiple of } \delta}} E_{\hat{\theta}|\tilde{\mu}_4}\left[\frac{\left|\hat{\theta}-\theta\right|}{\sigma}\right]$$

where $\tilde{\mu}_4$ is the kurtosis specifying the two-parameter distribution, $E_{\hat{\theta}|\tilde{\mu}_4}$ denotes the expected value given fixed $\tilde{\mu}_4$.

Standardization plays a crucial role in comparing the performance of estimators across different distributions. Currently, several options are available, such as using the root mean square deviation from the mode (as in Gauss (1)), the mean absolute deviation, or the standard deviation. However, the standard deviation is preferred due to its central role in standard error estimation. In Table 1, $\delta = 0.1$, $C = 70$. For the Weibull, gamma, lognormal and generalized Gaussian distributions, $\tilde{\mu}_{4_l} = 3$ (there are two shape parameter solutions for the Weibull distribution, the lower one is used here). For the Pareto distribution, $\tilde{\mu}_{4_l} = 9$. To provide a more practical and straightforward illustration, all results from five distributions are further weighted by the number of Google Scholar search results. Within the range of kurtosis setting, nearly all WLs and WHL**k**ms proposed here reach or at least come close to their maximum biases (SI Dataset S1). The pseudo-maximum bias is thus defined as the maximum value of the biases within the range of kurtosis setting for all five unimodal distributions. In most cases, the pseudo-maximum biases of invariant moments occur in lognormal or generalized Gaussian distributions (SI Dataset S1), since besides unimodality, the Weibull distribution differs entirely from them. Interestingly, the asymptotic biases of $\mathrm{TM}_{\epsilon=\frac{1}{24}}$ and $\mathrm{WM}_{\epsilon=\frac{1}{24}}$, after averaging and weighting, are 0.000σ and 0.000σ, respectively, in line with the sharp bias bounds of $\mathrm{TM}_{2,14:15}$ and $\mathrm{WM}_{2,14:15}$ (a different subscript is used to indicate a sample size of 15, with the removal of the first and last order statistics), 0.173σ and 0.126σ, for distributions with finite moments without assuming unimodality (48, 49).

Discussion

Moments, including raw moments, central moments, and standardized moments, are the most common parameters that describe probability distributions. Central moments are preferred over raw moments because they are invariant to translation. In 1947, Hsu and Robbins proved that the arithmetic mean converges completely to the population mean provided the second moment is finite (52). The strong law of large numbers (proven by Kolmogorov in 1933) (53) implies that the kth sample central moment is asymptotically unbiased. Recently, fascinating statistical phenomena regarding Taylor's law for distributions with infinite moments have been discovered by Drton and Xiao (2016) (54), Pillai and Meng (2016) (55), Cohen, Davis, and Samorodnitsky (2020) (56), and Brown, Cohen, Tang, and Yam (2021) (57). Lindquist and Rachev (2021) raised a critical question in their inspiring comment to Brown et al's paper (57): "What are the proper measures for the location, spread, asymmetry, and dependence (association) for random samples with infinite mean?" (58). From a different perspective, this question closely aligns with the essence of Bickel and Lehmann's open question in 1979 (10). They suggested using median, interquartile range, and medcouple (59) as the robust versions of the first three moments. While answering this question is not the focus of this paper, it is almost certain that the estimators proposed in this series will have a place. Since the efficiency of an L-statistic to the sample mean is generally monotonic with respect to the breakdown point (18), and the estimation of central moments can be transformed into the location estimation of the central moment kernel distribution, similar monotonic relations can be expected. In the case of a distribution with an infinite mean, non-robust estimators will not converge and will not provide valid estimates since their variances will be infinitely large. Therefore, the desired measures should be as robust as possible. Clearly now, if one wants to preserve the original relationship between each moment while ensuring maximum robustness, the natural choices are median, median variance, and median skewness. Similar to the robust version of L-moment (60) being trimmed L-moment (15), mean and central moments now also have their standard most robust version based on the complete congruence of the underlying distribution.

More generally, statistics, encompassing the collection, analysis, interpretation, and presentation of data, has evolved over time, with various approaches emerging to meet challenges in practice. Among these approaches, the use of probability models and measures of random variables for data analysis is often considered the core of statistics. While the early development of statistics was focused on parametric methods, there were two main approaches to point estimation. The Gauss–Markov theorem (1, 61) states the principle of minimum variance unbiased estimation which was further enriched by Neyman (1934) (62), Rao (1945) (63), Blackwell (1947) (64), and Lehmann and Scheffé (1950, 1955) (30, 31). Maximum likelihood was first introduced by Fisher in 1922 (65) in a multinomial model and later generalized by Cramér (1946), Hájek (1970), and Le Cam (1972) (40, 66, 67). In 1939, Wald (68) combined these two principles and suggested the use of minimax estimates, which involve choosing an estimator that minimizes the maximum possible loss. Hodges and Lehmann in 1950 (69) expanded upon this concept and obtained minimax estimates for a series of important problems. Following Huber's seminal work (3), M-statistics have dominated the field of parametric robust statistics for over half a century. Nonparametric methods, e.g., the Kolmogorov–Smirnov test, Mann-Whitney-Wilcoxon Test, and Hoeffding's independence test, emerged as popular alternatives to parametric methods in 1950s, as they do not make specific assumptions about the underlying distribution of the data. In 1963, Hodges and Lehmann proposed a class of robust location estimators based on the confidence bounds of rank tests (70). In RMSM I, when compared to other semiparametric mean estimators with the same breakdown point, the H-L estimator was shown to be the bias-optimal choice, which aligns Devroye, and Lerasle, Lugosi, and Oliveira's conclusion that the median of means is near-optimal in terms of concentration bounds (42) as discussed. The formal study of semiparametric models was initiated by Stein (71) in 1956. Bickel, in 1982, simplified the general heuristic necessary condition proposed by Stein (71) and derived sufficient conditions for this type of problem, adaptive estimation (72). These conditions were subsequently applied to the construction of adaptive estimates (72). It has become increasingly apparent that, in robust statistics, many estimators previously called "nonparametric" are essentially semiparametric as they are partly, though not fully, characterized by some interpretable Euclidean parameters. This approach is particularly useful in situations where the data do not conform to a simple parametric distribution but still have some structure that can be exploited. In 1984, Bickel addressed the challenge of robustly estimating the parameters of a linear model while acknowledging the possibility that the model may be invalid but still within the confines of a larger model (73). He showed by carefully designing the estimators, the biases can be very small. The paradigm shift here opens up the possibility that by defining a large semiparametric model and constructing estimators simultaneously for two or more very different semiparametric/parametric models within the large semiparametric model, then even for a parametric model belongs to the large semiparametric model but not to the semiparametric/parametric models used for calibration, the performance of these estimators might still be near-optimal due to the common nature shared by the models used by the estimators. Closely related topics are "mixture model" and "constraint defined model," which were generalized in Bickel, Klaassen, Ritov, and Wellner's classic semiparametric textbook (1993) (74) and the method of sieves, introduced by Grenander in 1981 (75). As the building blocks of statistics, invariant moments can improve the consistency of statistical results across studies, particularly when heavy-tailed distributions may be present (76, 77).

1. CF Gauss, *Theoria combinationis observationum erroribus minimis obnoxiae.* (Henricus Dieterich), (1823).
2. S Newcomb, A generalized theory of the combination of observations so as to obtain the best result. *Am. journal Math.* **8**, 343–366 (1886).
3. PJ Huber, Robust estimation of a location parameter. *Ann. Math. Stat.* **35**, 73–101 (1964).
4. X He, WK Fung, Method of medians for lifetime data with weibull models. *Stat. medicine* **18**, 1993–2009 (1999).
5. M Menon, Estimation of the shape and scale parameters of the weibull distribution. *Technometrics* **5**, 175–182 (1963).
6. SD Dubey, Some percentile estimators for weibull parameters. *Technometrics* **9**, 119–129 (1967).
7. NB Marks, Estimation of weibull parameters from common percentiles. *J. applied Stat.* **32**, 17–24 (2005).
8. K Boudt, D Caliskan, C Croux, Robust explicit estimators of weibull parameters. *Metrika* **73**, 187–209 (2011).
9. PJ Bickel, EL Lehmann, Descriptive statistics for nonparametric models. iii. dispersion in *Selected works of EL Lehmann.* (Springer), pp. 499–518 (2012).
10. PJ Bickel, EL Lehmann, Descriptive statistics for nonparametric models iv. spread in *Selected Works of EL Lehmann.* (Springer), pp. 519–526 (2012).
11. AL Bowley, *Elements of statistics.* (King) No. 8, (1926).
12. WR van Zwet, *Convex Transformations of Random Variables: Nebst Stellingen.* (1964).
13. RA Groeneveld, G Meeden, Measuring skewness and kurtosis. *J. Royal Stat. Soc. Ser. D (The Stat.* **33**, 391–399 (1984).
14. J SAW, Moments of sample moments of censored samples from a normal population. *Biometrika* **45**, 211–221 (1958).
15. EA Elamir, AH Seheult, Trimmed l-moments. *Comput. Stat. & Data Analysis* **43**, 299–314 (2003).
16. H Oja, On location, scale, skewness and kurtosis of univariate distributions. *Scand. J. statistics* pp. 154–168 (1981).
17. H Oja, Descriptive statistics for multivariate distributions. *Stat. & Probab. Lett.* **1**, 327–332 (1983).
18. PJ Bickel, EL Lehmann, Descriptive statistics for nonparametric models ii. location in *selected works of EL Lehmann.* (Springer), pp. 473–497 (2012).
19. W van Zwet, Convex transformations: A new approach to skewness and kurtosis in *Selected Works of Willem van Zwet.* (Springer), pp. 3–11 (2012).
20. PJ Rousseeuw, C Croux, Alternatives to the median absolute deviation. *J. Am. Stat. association* **88**, 1273–1283 (1993).
21. PM Heffernan, Unbiased estimation of central moments by using u-statistics. *J. Royal Stat. Soc. Ser. B (Statistical Methodol.* **59**, 861–863 (1997).
22. J Hodges, E Lehmann, Matching in paired comparisons. *The Annals Math. Stat.* **25**, 787–791 (1954).
23. RA Fisher, Moments and product moments of sampling distributions. *Proc. Lond. Math. Soc.* **2**, 199–238 (1930).
24. PR Halmos, The theory of unbiased estimation. *The Annals Math. Stat.* **17**, 34–43 (1946).
25. W Hoeffding, A class of statistics with asymptotically normal distribution. *The Annals Math. Stat.* **19**, 293–325 (1948).
26. RJ Serfling, Generalized l-, m-, and r-statistics. *The Annals Stat.* **12**, 76–86 (1984).
27. E Joly, G Lugosi, Robust estimation of u-statistics. *Stoch. Process. their Appl.* **126**, 3760–3773 (2016).
28. P Laforgue, S Clémençon, P Bertail, On medians of (randomized) pairwise means in *International Conference on Machine Learning.* (PMLR), pp. 1272–1281 (2019).
29. P Bickel, E Lehmann, Descriptive statistics for nonparametric models i. introduction in *Selected Works of EL Lehmann.* (Springer), pp. 465–471 (2012).
30. EL Lehmann, H Scheffé, Completeness, similar regions, and unbiased estimation-part i in *Selected works of EL Lehmann.* (Springer), pp. 233–268 (2011).
31. EL Lehmann, H Scheffé, *Completeness, similar regions, and unbiased estimation—part II.* (Springer), (2012).
32. T Lai, H Robbins, K Yu, Adaptive choice of mean or median in estimating the center of a symmetric distribution. *Proc. Natl. Acad. Sci.* **80**, 5803–5806 (1983).
33. PS Laplace, *Theorie analytique des probabilities.* (1812).
34. B Efron, Bootstrap methods: Another look at the jackknife. *The Annals Stat.* **7**, 1–26 (1979).
35. PJ Bickel, DA Freedman, Some asymptotic theory for the bootstrap. *The annals statistics* **9**, 1196–1217 (1981).
36. R Helmers, P Janssen, N Veraverbeke, *Bootstrapping U-quantiles.* (CWI. Department of Operations Research, Statistics, and System Theory [BS]), (1990).
37. RD Richtmyer, A non-random sampling method, based on congruences, for" monte carlo" problems, (New York Univ., New York. Atomic Energy Commission Computing and Applied . . .), Technical report (1958).
38. IM Sobol', On the distribution of points in a cube and the approximate evaluation of integrals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki* **7**, 784–802 (1967).
39. KA Do, P Hall, Quasi-random resampling for the bootstrap. *Stat. Comput.* **1**, 13–22 (1991).
40. H Cramér, *Mathematical methods of statistics.* (Princeton university press) Vol. 43, (1999).
41. FR Hampel, *Contributions to the theory of robust estimation.* (University of California, Berkeley), (1968).
42. L Devroye, M Lerasle, G Lugosi, RI Oliveira, Sub-gaussian mean estimators. *The Annals Stat.* **44**, 2695–2725 (2016).
43. SJ Devlin, R Gnanadesikan, JR Kettenring, Robust estimation and outlier detection with correlation coefficients. *Biometrika* **62**, 531–545 (1975).
44. FR Hampel, A general qualitative definition of robustness. *The annals mathematical statistics* **42**, 1887–1896 (1971).
45. FR Hampel, The influence curve and its role in robust estimation. *J. american statistical association* **69**, 383–393 (1974).
46. PJ Rousseeuw, FR Hampel, EM Ronchetti, WA Stahel, *Robust statistics: the approach based on influence functions.* (John Wiley & Sons), (2011).
47. DL Donoho, PJ Huber, The notion of breakdown point. *A festschrift for Erich L. Lehmann* **157184** (1983).
48. M Bieniek, Comparison of the bias of trimmed and winsorized means. *Commun. Stat. Methods* **45**, 6641–6650 (2016).
49. K Danielak, T Rychlik, Theory & methods: Exact bounds for the bias of trimmed means. *Aust. & New Zealand J. Stat.* **45**, 83–96 (2003).
50. C Bernard, R Kazzi, S Vanduffel, Range value-at-risk bounds for unimodal distributions under partial information. *Insur. Math. Econ.* **94**, 9–24 (2020).
51. T Mathieu, Concentration study of m-estimators using the influence function. *Electron. J. Stat.* **16**, 3695–3750 (2022).
52. PL Hsu, H Robbins, Complete convergence and the law of large numbers. *Proc. national academy sciences* **33**, 25–31 (1947).
53. A Kolmogorov, Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.* **4**, 83–91 (1933).
54. M Drton, H Xiao, Wald tests of singular hypotheses. *Bernoulli* **22**, 38–59 (2016).
55. NS Pillai, XL Meng, An unexpected encounter with cauchy and lévy. *The Annals Stat.* **44**, 2089–2097 (2016).
56. JE Cohen, RA Davis, G Samorodnitsky, Heavy-tailed distributions, correlations, kurtosis and taylor's law of fluctuation scaling. *Proc. Royal Soc. A* **476**, 20200610 (2020).
57. M Brown, JE Cohen, CF Tang, SCP Yam, Taylor's law of fluctuation scaling for semivariances and higher moments of heavy-tailed data. *Proc. Natl. Acad. Sci.* **118**, e2108031118 (2021).
58. WB Lindquist, ST Rachev, Taylor's law and heavy-tailed distributions. *Proc. Natl. Acad. Sci.* **118**, e2118893118 (2021).
59. G Brys, M Hubert, A Struyf, A robust measure of skewness. *J. Comput. Graph. Stat.* **13**, 996–1017 (2004).
60. JR Hosking, L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *J. Royal Stat. Soc. Ser. B (Methodological)* **52**, 105–124 (1990).
61. AA Markov, *Wahrscheinlichkeitsrechnung.* (Teubner), (1912).
62. J Neyman, On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *J. Royal Stat. Soc.* **97**, 558–606 (1934).
63. C Radhakrishna Rao, Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **37**, 81–91 (1945).
64. D Blackwell, Conditional expectation and unbiased sequential estimation. *The Annals Math. Stat.* pp. 105–110 (1947).
65. RA Fisher, On the mathematical foundations of theoretical statistics. *Philos. transactions Royal Soc. London. Ser. A, containing papers a mathematical or physical character* **222**, 309–368 (1922).
66. L LeCam, On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *The Annals Math. Stat.* **41**, 802–828 (1970).
67. J Hájek, Local asymptotic minimax and admissibility in estimation in *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability*. Vol. 1, pp. 175–194 (1972).
68. A Wald, Contributions to the theory of statistical estimation and testing hypotheses. *The Annals Math. Stat.* **10**, 299–326 (1939).
69. J Hodges, EL Lehmann, Some problems in minimax point estimation in *Selected Works of EL Lehmann.* (Springer), pp. 15–30 (2012).
70. J Hodges Jr, E Lehmann, Estimates of location based on rank tests. *The Annals Math. Stat.* **34**, 598–611 (1963).
71. CM Stein, Efficient nonparametric testing and estimation in *Proceedings of the third Berkeley symposium on mathematical statistics and probability*. Vol. 1, pp. 187–195 (1956).
72. PJ Bickel, On adaptive estimation. *The Annals Stat.* **10**, 647–671 (1982).
73. P Bickel, Parametric robustness: small biases can be worthwhile. *The Annals Stat.* **12**, 864–879 (1984).
74. P Bickel, CA Klaassen, Y Ritov, JA Wellner, *Efficient and adaptive estimation for semiparametric models.* (Springer) Vol. 4, (1993).
75. U Grenander, *Abstract Inference.* (1981).
76. JT Leek, RD Peng, Reproducible research can still be wrong: adopting a prevention approach. *Proc. Natl. Acad. Sci.* **112**, 1645–1646 (2015).
77. E National Academies of Sciences, et al., *Reproducibility and Replicability in Science.* (National Academies Press), (2019).