Near-consistent robust estimations of moments for unimodal distributions

Tuban Lee[a,1]

[a] Institute of Biomathematics, Macau SAR 999078, China

This manuscript was compiled on March 25, 2023

Descriptive statistics for parametric models currently heavily rely on the accuracy of distributional assumptions. Here, based on the invariant structures of unimodal distributions, a series of sophisticated yet efficient estimators, robust to both gross errors and departures from parametric assumptions, are proposed for estimating mean and central moments with insignificant asymptotic biases for common unimodal distributions. This article also illuminates the understanding of the common nature of probability distributions and the measures of them.

orderliness | invariant | unimodal | adaptive estimation | U-statistics

The asymptotic inconsistencies between sample mean ($\bar{x}$) and nonparametric robust location estimators in asymmetric distributions on the real line have been noticed for more than two centuries (1), yet remain unsolved. Strictly speaking, it is unsolvable as by trimming, some information about the original distribution is removed, making it impossible to estimate the values of the removed parts without distributional assumptions. Newcomb (1886, 1912) 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, 3). In 1964, Huber (4) 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 parameter estimations. For example, He and Fung (1999) constructed (5) a robust M-estimator for the two-parameter Weibull distribution, from which all moments can be calculated. Nonetheless, it is inadequate for the gamma, Perato, lognormal, and the generalized Gaussian distributions (SI Dataset S1). Another old and interesting approach is arithmetically computing the parameters using one or more L-statistics as inputs, such as percentile estimators. Examples of percentile estimators for the Weibull distribution, the reader is referred to Menon (1963) (6), Dubey (1967) (7), Hassanein (1971) (8), Marks (2005) (9), and Boudt, Caliskan, and Croux (2011) (10)'s works. At the outset of the study of percentile estimators, it was known that they arithmetically utilizes the invariant structures of probability distributions (6, 11, 12). Maybe such estimators can be named as I-statistics. Formally, an estimator is classified as an I-statistic if it asymptotically satisfies $\mathrm{I}\left(\mathrm{LE}_1, \cdots, \mathrm{LE}_l\right) = \left(\theta_1, \cdots, \theta_q\right)$ for the distribution it is consistent, where LEs are calculated with the use of L-statistics, I is defined using arithmetic operations and constants, but it may also incorporate other functions, and θs are the population parameters it estimates. A subclass of I-statistics, arithmetic I-statistics, is defined as LEs are L-statistics, I is solely defined using arithmetic operations and constants. Since some percentile estimators use the logarithmic function to transform all random variables before computing the L-statistics, a percentile estimator might not always be an arithmetic I-statistic (7). In this article, two subclasses of I-statistics are introduced, arithmetic I-statistics and quantile I-statistics. Examples of quantile I-statistics will be discussed later. Based on L-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-statistic can be expressed as an integral of the quantile function, I-statistics are often analytically obtainable. However, the performance of the aforementioned examples is often worse than that of the robust M-statistics when the distributional assumption is violated (SI Dataset S1). Even when distributions such as the Weibull and gamma belong to the same larger family, the generalized gamma distribution, a misassumption can still result in substantial biases, rendering the approach ill-suited.

In previous research on semiparametric robust mean estimation, the binomial Hodges-Lehmann mean (BHLM_ϵ) is still inconsistent for any skewed distribution, despite having much smaller asymptotic biases than other symmetric weighted L-statistics, which are either symmetric weighted averages or symmetric weighted H-L means. All robust location estimators commonly used are symmetric due to the universality of the symmetric distributions. One can construct an asymmetric weighted average that is consistent for a semiparametric class of skewed distributions. This approach has been investigated previously, but its lack of symmetry makes it suitable only for certain applications (13). Shifting from semiparametrics to parametrics, an ideal robust location estimator would have a non-sample-dependent breakdown point (defined in Subsection F) and be consistent for any symmetric distribution and a

Significance Statement

Bias, variance, and contamination are the three main errors in statistics. Consistent robust estimation is unattainable without parametric assumptions. Here, based on a paradigm shift inspired by mean-median-mode inequality, Bickel-Lehmann spread, and adaptive estimation, 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

skewed distribution with finite second moments. This is called an invariant mean. Based on the mean-symmetric weighted L-statistic-median inequality, the recombined mean is defined as

$$rm_{d,\epsilon,n} \coloneqq \lim_{c\to\infty}\left(\frac{(\mathrm{SWL}_{\epsilon,n}+c)^{d+1}}{(m_n+c)^d}-c\right),$$

where d is the key factor for bias correction, m_n is the sample median, $\mathrm{SWL}_{\epsilon,n}$ is $\mathrm{BM}_{\epsilon,n}$ in the first Subsection, but other symmetric weighted L-statistics can also be used in practice as long as the inequalities hold. The following theorem shows the significance of this arithmetic I-statistic.

Theorem .1. *If the second moments are finite, $rm_{d\approx 0.375,\epsilon=\frac{1}{8}}$ is a consistent mean estimator for the exponential and any symmetric distributions and the Pareto distribution with quantile function $Q(p)=x_m(1-p)^{-\frac{1}{\alpha}}$, $x_m>0$, when $\alpha\to\infty$.*

Proof. Finding d and ϵ that make $rm_{d,\epsilon}$ a consistent mean estimator is equivalent to finding the solution of $E[rm_{d,\epsilon,n}]=E[X]$. Rearranging the definition, $rm_{d,\epsilon}=\lim_{c\to\infty}\left(\frac{(\mathrm{BM}_\epsilon+c)^{d+1}}{(m+c)^d}-c\right)=(d+1)\,\mathrm{BM}_\epsilon-dm=\mu$. So, $d=\frac{\mu-\mathrm{BM}_\epsilon}{\mathrm{BM}_\epsilon-m}$. The quantile function of the exponential distribution is $Q(p)=\ln\left(\frac{1}{1-p}\right)\lambda$. $E[X]=\lambda$. $E[m_n]=Q\left(\frac{1}{2}\right)=\ln 2\lambda$. For the exponential distribution, $E\left[\mathrm{BM}_{\frac{1}{8},n}\right]=\lambda\left(1+\ln\left(\frac{46656}{8575\sqrt{35}}\right)\right)$. Obviously, the scale parameter λ can be canceled out, $d\approx 0.375$. The proof of the second assertion follows directly from the coincidence property. For any symmetric distribution with a finite second moment, $E[\mathrm{BM}_{\epsilon,n}]=E[m_n]=E[X]$. Then $E[rm_{d,\epsilon,n}]=\lim_{c\to\infty}\left(\frac{(E[X]+c)^{d+1}}{(E[X]+c)^d}-c\right)=E[X]$. The proof for the Pareto distribution is more general. The mean of the Pareto distribution is given by $\frac{\alpha x_m}{\alpha-1}$. The d value with two unknown percentiles p_1 and p_2 for the Pareto distribution is $d_{Perato}=\frac{\frac{\alpha x_m}{\alpha-1}-x_m(1-p_1)^{-\frac{1}{\alpha}}}{x_m(1-p_1)^{-\frac{1}{\alpha}}-x_m(1-p_2)^{-\frac{1}{\alpha}}}$. Since any weighted L-statistic can be expressed as an integral of the quantile function, $\lim_{\alpha\to\infty}\frac{\frac{\alpha}{\alpha-1}-(1-p_1)^{-1/\alpha}}{(1-p_1)^{-1/\alpha}-(1-p_2)^{-1/\alpha}}=-\frac{\ln(1-p_1)+1}{\ln(1-p_1)-\ln(1-p_2)}$, the d value for the Pareto distribution approaches that of the exponential distribution as $\alpha\to\infty$, regardless of the type of weighted L-statistic used. This completes the demonstration. ☐

Theorem .1 implies that for the Weibull, gamma, Pareto, lognormal and generalized Gaussian distribution, $rm_{d\approx 0.375,\epsilon=\frac{1}{8}}$ is consistent for at least one particular case. The biases of $rm_{d\approx 0.375,\epsilon=\frac{1}{8}}$ for distributions with skewness between those of the exponential and symmetric distributions are tiny (SI Dataset S1). $rm_{d\approx 0.375,\epsilon=\frac{1}{8}}$ exhibits excellent performance for all these common unimodal distributions (SI Dataset S1).

Besides introducing the concept of invariant mean, 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 a series of sophisticated yet efficient robust estimators can be constructed whose biases are typically smaller than the variances (as seen in Table **??** for $n=5400$) for unimodal distributions.

Background and Main Results

A. Invariant mean. It is well established that a theoretical model can be adjusted to fit the first two moments of the observed data. A continuous distribution belonging to a location–scale family takes the form $F(x)=F_0\left(\frac{x-\mu}{\lambda}\right)$, where F_0 is a "standard" distribution. Therefore, $F(x)=Q^{-1}(x)\to x=Q(p)=\lambda Q_0(p)+\mu$. Thus, any L-statistic can be expressed as $\lambda \mathrm{L}_0(\epsilon)+\mu$, where $\mathrm{L}_0(\epsilon)$ is an integral of $Q_0(p)$ according to the definition of the L-statistic. The simultaneous cancellation of μ and λ in $\frac{(\lambda\mu_0+\mu)-(\lambda \mathrm{L}_0(\epsilon)+\mu)}{(\lambda \mathrm{L}_0(\epsilon)+\mu)-(\lambda m_0+\mu)}$ assures that d is a constant. Consequently, the roles of SWL_ϵ and median in $rm_{d,\epsilon}$ can be replaced by any L-statistics, although only symmetric weighted L-statistics are considered in defining the invariant mean.

The performance in heavy-tailed distributions can be further improved by constructing the quantile mean as

$$qm_{d,\epsilon,n} \coloneqq \hat{Q}_n\left(\left(\hat{F}_n(\mathrm{SWL}_{\epsilon,n})-\frac{1}{2}\right)d+\hat{F}_n(\mathrm{SWL}_{\epsilon,n})\right),$$

provided that $\hat{F}_n(\mathrm{SWL}_{\epsilon,n})\geq\frac{1}{2}$, where $\hat{F}_n(x)$ is the empirical cumulative distribution function of the sample, $\hat{Q}_n$ is the sample quantile function. The most popular method for computing the sample quantile function was proposed by Hyndman and Fan in 1996 (14). To minimize the finite sample bias, here, $\hat{F}_n(x)\coloneqq\frac{1}{n}\left(\frac{x-X_{sp}}{X_{sp+1}-X_{sp}}+sp\right)$, where $sp=\sum_{i=1}^{n}\mathbf{1}_{X_i\leq x}$, $\mathbf{1}_A$ is the indicator of event A. The solution of $\hat{F}_n(\mathrm{SWL}_{\epsilon,n})<\frac{1}{2}$ is reversing the percentile by $1-\hat{F}_n(\mathrm{SWL}_{\epsilon,n})$, the obtained percentile is also reversed. Without loss of generality, in the following discussion, only the case where $\hat{F}_n(\mathrm{SWL}_{\epsilon,n})\geq\frac{1}{2}$ is considered. Moreover, in extreme heavy-tailed distributions, the calculated percentile can exceed the breakdown point of SWL_ϵ, so the percentile will be modified to $1-\epsilon$ if this occurs. The quantile mean uses the location-scale invariant in a different way as shown in the following proof.

Theorem A.1. *$qm_{d\approx 0.321,\epsilon=\frac{1}{8}}$ is a consistent mean estimator for the exponential, Pareto ($\alpha\to\infty$) and any symmetric distributions provided that the second moments are finite.*

Proof. Similarly, rearranging the definition, $d=\frac{F(\mu)-F(\mathrm{BM}_\epsilon)}{F(\mathrm{BM}_\epsilon)-\frac{1}{2}}$. The cdf of the exponential distribution is $F(x)=1-e^{-\lambda^{-1}x}$, $\lambda\geq 0$, $x\geq 0$, the expectation of $\mathrm{BM}_{\epsilon,n}$ can be expressed as $\lambda\mathrm{BM}_0(\epsilon)$, so $F(\mathrm{BM}_\epsilon)$ is free of λ. When $\epsilon=\frac{1}{8}$, $d=\frac{-e^{-1}+e^{-\left(1+\ln\left(\frac{46656}{8575\sqrt{35}}\right)\right)}}{\frac{1}{2}-e^{-\left(1+\ln\left(\frac{46656}{8575\sqrt{35}}\right)\right)}}\approx 0.321$. The proof of the symmetric case is similar. Since for any symmetric distribution with a finite second moment, $F(E[\mathrm{BM}_{\epsilon,n}])=F(\mu)=\frac{1}{2}$. Then, the expectation of the quantile mean is $qm_{d,\epsilon}=F^{-1}\left(\left(F(\mu)-\frac{1}{2}\right)d+F(\mu)\right)=F^{-1}(0+F(\mu))=\mu$.

For the assertion related to the Pareto distribution, the cdf of it is $1-\left(\frac{x_m}{x}\right)^\alpha$. So, the d value with two unknown percentile p_1 and p_2 is

$$d_{Pareto}=\frac{1-\left(\frac{x_m}{\frac{\alpha x_m}{\alpha-1}}\right)^\alpha-\left(1-\left(\frac{x_m}{x_m(1-p_1)^{-\frac{1}{\alpha}}}\right)^\alpha\right)}{\left(1-\left(\frac{x_m}{x_m(1-p_1)^{-\frac{1}{\alpha}}}\right)^\alpha\right)-\left(1-\left(\frac{x_m}{x_m(1-p_2)^{-\frac{1}{\alpha}}}\right)^\alpha\right)}=$$

$\frac{1-\left(\frac{\alpha-1}{\alpha}\right)^{\alpha}-p_1}{p_1-p_2}$. When $\alpha \to \infty$, $\left(\frac{\alpha-1}{\alpha}\right)^{\alpha} = \frac{1}{e}$. The d value for the exponential distribution is identical, since $d_{exp} = \frac{\left(1-e^{-1}\right)-\left(1-e^{-\ln\left(\frac{1}{1-p_1}\right)}\right)}{\left(1-e^{-\ln\left(\frac{1}{1-p_1}\right)}\right)-\left(1-e^{-\ln\left(\frac{1}{1-p_2}\right)}\right)} = \frac{1-\frac{1}{e}-p_1}{p_1-p_2}$. All results are now proven. □

The definitions of location and scale parameters are such that they must satisfy $F(x;\lambda,\mu) = F(\frac{x-\mu}{\lambda};1,0)$. By recalling $x = \lambda Q_0(p) + \mu$, it follows that the percentile of any weighted L-statistic is free of λ and μ, which guarantees the validity of the quantile mean. The quantile mean is a quantile I-statistic. Specifically, an estimator is classified as a quantile I-statistic if LEs are percentiles of a distribution obtained by plugging L-statistics into a cumulative distribution function and I is defined with arithmetic operations, constants and quantile functions. $qm_{d\approx 0.321,\epsilon=\frac{1}{8}}$ works better in the fat-tail scenarios (SI Dataset S1). Theorem .1 and A.1 show that $rm_{d\approx 0.375,\epsilon=\frac{1}{8}}$ and $qm_{d\approx 0.321,\epsilon=\frac{1}{8}}$ are both consistent mean estimators for any symmetric distribution and a skewed distribution with finite second moments. It's obvious that the breakdown points of $rm_{d\approx 0.375,\epsilon=\frac{1}{8}}$ and $qm_{d\approx 0.321,\epsilon=\frac{1}{8}}$ are both $\frac{1}{8}$. Therefore they are all invariant means.

To study the impact of the choice of SWAs in rm and qm, it is constructive to recall that a symmetric weighted average is a linear combination of symmetric quantile averages. While using a less-biased symmetric weighted average can generally enhance performance (SI Dataset S1), there is a greater risk of violation in the semiparametric framework. However, the mean-SWA$_\epsilon$-median inequality is robust to slight fluctuations of the SQA function of the underlying distribution. Suppose the SQA function is generally decreasing in $[0,u]$, but increasing in $[u,\frac{1}{2}]$, since $1-2\epsilon$ of the symmetric quantile averages will be included in the computation of SWA$_\epsilon$, as long as $\frac{1}{2}-u \ll 1-2\epsilon$, and other portions of the SQA function satisfy the inequality constraints that define the νth orderliness on which the SWA$_\epsilon$ is based, the mean-SWA$_\epsilon$-median inequality will still hold. This is due to the violation being bounded (15) and therefore cannot be extreme for unimodal distributions. For instance, the SQA function is non-monotonic when the shape parameter of the Weibull distribution $\alpha > \frac{1}{1-\ln(2)} \approx 3.259$ as shown in the previous article, the violation of the third orderliness starts near this parameter as well, yet the mean-BM$_{\frac{1}{8}}$-median inequality is still valid when $\alpha \leq 3.322$. Another key factor in determining the risk of violation is the skewness of the distribution. Previously, 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, often only occurs when the distribution is nearly symmetrical (16). The over-corrections in rm and qm are dependent on the SWA$_\epsilon$-median difference, which can be a reasonable measure of skewness (17, 18), implying that the over-correction is often tiny with a moderate d. The sample logic can be applied to SWHLM. This qualitative analysis provides another perspective, in addition to the bias bounds (15), that rm and qm based on the mean-SWL$_\epsilon$-median inequality are generally safe.

B. Robust estimations of the central moments. In 1979, Bickel and Lehmann, in their final paper of the landmark series *Descriptive Statistics for Nonparametric Models* (19), generalized a class of estimators called "measures of spread," which "does not require the assumption of symmetry." From that, a popular efficient scale estimator, the Rousseeuw-Croux scale estimator (20), was derived in 1993, but the importance of tackling the symmetry assumption has been greatly underestimated. While they had already considered one version of the trimmed standard deviation in the third paper of that series (21), in the final section of that paper (19), they explored another two possible versions, which were modified here for comparison,

$$\left[n\left(\frac{1}{2}-\epsilon\right)\right]^{-\frac{1}{2}}\left[\sum_{i=\frac{n}{2}}^{n(1-\epsilon)}\left[X_i-X_{n-i+1}\right]^2\right]^{\frac{1}{2}}, \qquad [1]$$

and

$$\left[\binom{n}{2}(1-\epsilon-\gamma\epsilon)\right]^{-\frac{1}{2}}\left[\sum_{i=\binom{n}{2}\gamma\epsilon}^{\binom{n}{2}(1-\epsilon)}\left(X-X'\right)_i^2\right]^{\frac{1}{2}}, \qquad [2]$$

where $(X-X')_1 \leq \ldots \leq (X-X')_{\binom{n}{2}}$ are the order statistics of the "pseudo-sample", $X_i - X_j$, $i<j$. The paper ended with, "We do not know a fortiori which of the measures [1] or [2] is preferable and leave these interesting questions open."

Observe that the kernel of the unbiased estimation of the second central moment by using U-statistic is $\psi_2(x_1,x_2) = \frac{1}{2}(x_1-x_2)^2$. If adding the $\frac{1}{2}$ term in [2], as $\epsilon \to 0$, the result is equivalent to the standard deviation estimated by using U-statistic (also noted by Janssen, Serfling, and Veraverbeke in 1987) (22). In fact, they also showed that, when ϵ is 0, [2] is $\sqrt{2}$ times the standard deviation.

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

Nomenclature. Given a robust estimator $\hat{\theta}$ with an adjustable breakdown point which can be infinitesimal, the name of $\hat{\theta}$ is composed of two parts: the first part denotes the type of estimator, and the second part is the name of the population parameter θ that the estimator is consistent with as $\epsilon \to 0$. The abbreviation of the estimator is formed by combining the initial letter(s) of the first part with the common abbreviation of the consistent estimator that measures the population parameter. If the estimator is symmetric, the asymptotic breakdown point, ϵ (or ϵ_U, if the estimator is a U-statistic), is indicated in the subscript of the abbreviation of the estimator, except the median. For an asymmetric estimator based on quantile average, the corresponding γ is also indicated after ϵ, the upper breakdown point (defined in Subsection F).

In the previous article on semiparametric robust mean estimation, 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 correspond to the population parameter with which it is consistent as $\epsilon \to 0$. The trimmed standard deviation following this nomenclature is $\mathrm{T}sd_{\epsilon_{U_2}=1-\sqrt{1-\epsilon},\gamma,n} := \left[\mathrm{TM}_{\epsilon,\gamma}\left(\left(\psi_2\left(X_{N_1},X_{N_2}\right)\right)_{N=1}^{\binom{n}{2}}\right)\right]^{-\frac{1}{2}}$, where $\mathrm{TM}_{\epsilon,\gamma}(Y)$ denotes the ϵ,γ-trimmed mean with the sequence $\left(\psi_2\left(X_{N_1},X_{N_2}\right)\right)_{N=1}^{\binom{n}{2}}$ as an input, the proof of the breakdown point is given in Subsection F. Removing the square root yields the trimmed variance ($\mathrm{T}var_{\epsilon_{U_2},\gamma,n}$). It is now very clear that this definition, essentially the same as [2], should be preferable. Not only because it is essentially a trimmed U-statistic for the standard deviation but also because the γ-orderliness of the second central moment kernel distribution is ensured by the next exciting theorem.

Theorem B.1. *The second central moment kernel distribution generated from any unimodal distribution is γ-ordered.*

Proof. The monotonic increasing of the pairwise difference distribution was first implied in its unimodality proof done by Hodges and Lehmann in 1954 (23). Whereas they used induction to get the result in Theorem **??**, Dharmadhikari and Jogdeo in 1982 (24) provided a modern proof of the unimodality using Khintchine's representation (25). Assuming absolute continuity, Purkayastha (26) introduced a much simpler proof in 1998. Transforming the pairwise difference distribution by squaring and multiplying by $\frac{1}{2}$ does not change the monotonicity, making the pdf become monotonically decreasing with mode at zero. In the previous article, it was proven that a right skewed distribution with a monotonic decreasing pdf is always γ-ordered, which gives the desired result. □

Previously, it was shown that any symmetric distribution with a finite second moment is νth ordered, indicating that orderliness does not require unimodality, e.g., a symmetric bimodal distribution is also ordered. An analysis of the Weibull distribution showed that unimodality does not guarantee orderliness. Theorem B.1 reveals another profound relationship between unimodality and orderliness, which is sufficient for trimming inequality.

In 1928, Fisher constructed k-statistics as unbiased estimators of cumulants (27). Halmos (1946) proved that the functional θ admits an unbiased estimator if and only if it is a regular statistical functional of degree k and showed a relation of symmetry, unbiasness and minimum variance (28). In 1948, Hoeffding generalized U-statistics (29) which enable the derivation of a minimum-variance unbiased estimator from each unbiased estimator of an estimable parameter. Heffernan (1997) (30) obtained an unbiased estimator of the kth central moment by using U-statistics and demonstrated that it is the minimum variance unbiased estimator for distributions with finite moments (31, 32). 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 U-statistic structure (33). Also in 1984, Janssen and Serfling and Veraverbeke (34) showed that the Bickel-Lehmann spread also belongs to the same class. It gradually became clear that the Hodges-Lehmenn estimator and trimmed standard deviation are all trimmed U-statistics (35–37).

Extending the trimmed U-statistic to weighted U-statistic, i.e., replacing the trimmed mean with weighted L-statistic. The weighted kth central moment ($k \leq n$) is defined as,

$$\mathrm{W}km_{\epsilon_{U_k},\gamma,n} \coloneqq \mathrm{WL}_{\epsilon,\gamma,n}\left(\left(\psi_k\left(X_{N_1},\cdots,X_{N_k}\right)\right)_{N=1}^{\binom{n}{k}}\right),$$

where $\epsilon_{U_k} = 1 - (1-\epsilon)^{\frac{1}{k}}$, $X_{N_1},\cdots,X_{N_k}$ are the n choose k elements from X, $\psi_k\left(x_1,\cdots,x_k\right) = \sum_{j=0}^{k-2}\left(-1\right)^j\left(\frac{1}{k-j}\right)\sum\left(x_{i_1}^{k-j}\ldots x_{i_{(j+1)}}\right) + \left(-1\right)^{k-1}\left(k-1\right)x_1\ldots x_k$, the second summation is over $i_1,\cdots,i_{j+1} = 1$ to k with $i_1 < \cdots < i_{j+1}$ (30). Despite the complexity, the structure of the kth central moment kernel distributions can be elucidated by decomposing.

Theorem B.2. *For each pair $\left(Q\left(p_i\right),Q\left(p_j\right)\right)$ of the original distribution such that $Q\left(p_i\right) < Q\left(p_j\right)$, let $x_1 = Q(p_i)$ and $x_k = Q(p_j)$, $\Delta = Q(p_i) - Q(p_j)$, the kth central moment kernel distribution, $k > 2$, can be seen as a mixture distribution and each of the components has the support $\left(-\left(\frac{k}{\frac{3+(-1)^k}{2}}\right)^{-1}(-\Delta)^k, \frac{1}{k}(-\Delta)^k\right)$.*

Proof. Without loss of generality, generating the distribution of the k-tuple $(Q(p_{i_1}),\ldots,Q(p_{i_k}))$ under continuity, $k > 2$, $i_1 < \ldots < i_k$, $p_{i_1} < \ldots < p_{i_k}$, the corresponding probability density is $f_{X,\ldots,X}(Q(p_{i_1}),\ldots,Q(p_{i_k})) = k!f(Q(p_{i_1}))\ldots f(Q(p_{i_k}))$. Transforming the distribution of the k-tuple by the function $\psi_k\left(x_1,\ldots,x_k\right)$, denoting $\bar{\Delta} = \psi_k\left(Q(p_{i_1}),\ldots,Q(p_{i_k})\right)$. The probability $f_{\Xi_k}(\bar{\Delta}) = \sum_{\bar{\Delta}=\psi_k\left(Q(p_{i_1}),\ldots,Q(p_{i_k})\right)} f_{X,\ldots,X}(Q(p_{i_1}),\ldots,Q(p_{i_k}))$ is the summation of the probabilities of all k-tuples such that $\bar{\Delta}$ is equal to $\psi_k\left(Q(p_{i_1}),\ldots,Q(p_{i_k})\right)$. The following Ξ_k is equivalent.

Ξ_k: Every pair with a difference equal to $\Delta = Q(p_{i_1}) - Q(p_{i_k})$ can generate a pseudodistribution (but the integral is not equal to 1, so "pseudo") such that x_2, …, x_{k-1} exhaust all combinations under the inequality constraints, i.e., $Q(p_{i_1}) = x_1 < x_2 < \ldots < x_{k-1} < x_k = Q(p_{i_k})$. The combination of all the pseudodistributions with the same Δ is ξ_Δ. The combination of ξ_Δ, i.e., from $\Delta = 0$ to $Q(0) - Q(1)$, is Ξ_k.

The support of ξ_Δ is the extrema of ψ_k subject to the inequality constraints. Using the Lagrange multiplier, one can easily determine the only critical point at $x_1 = \ldots = x_k = 0$, where $\psi_k = 0$. Other candidates are within the boundaries, i.e., $\psi_k\left(x_1 = x_1, x_2 = x_k, \cdots, x_k = x_k\right)$, $\cdots$, $\psi_k\left(x_1 = x_1, \cdots, x_i = x_1, x_{i+1} = x_k, \cdots, x_k = x_k\right)$, $\cdots$, $\psi_k\left(x_1 = x_1, \cdots, x_{k-1} = x_1, x_k = x_k\right)$. $\psi_k\left(x_1 = x_1, \cdots, x_i = x_1, x_{i+1} = x_k, \cdots, x_k = x_k\right)$ can be divided into k groups. If $\frac{k+1-i}{2} \leq j \leq \frac{k-1}{2}$, from $j+1$st to $k-j$th group, the gth group has $i\binom{i-1}{g-j-1}\binom{k-i}{j}$ terms having the form $\left(-1\right)^{g+1}\frac{1}{k-g+1}x_1^{k-j}x_k^j$, from $k-j+1$th to $i+j$th group, the gth group has $i\binom{i-1}{g-j-1}\binom{k-i}{j} + \left(k-i\right)\binom{k-i-1}{j-k+g-1}\binom{i}{k-j}$ terms having the form $\left(-1\right)^{g+1}\frac{1}{k-g+1}x_1^{k-j}x_k^j$. If $j < \frac{k+1-i}{2}$, from $j+1$st to $i+j$th group, the gth group has $i\binom{i-1}{g-j-1}\binom{k-i}{j}$ terms having the form $\left(-1\right)^{g+1}\frac{1}{k-g+1}x_1^{k-j}x_k^j$. If $j \geq \frac{k}{2}$, from $k-j+1$st to jth group, the gth group has $\left(k-i\right)\binom{k-i-1}{j-k+g-1}\binom{i}{k-j}$ terms having the form $\left(-1\right)^{g+1}\frac{1}{k-g+1}x_1^{k-j}x_k^j$, from $j+1$th to $j+i$th group, $i+j < k$, the gth group has $i\binom{i-1}{g-j-1}\binom{k-i}{j} + \left(k-i\right)\binom{k-i-1}{j-k+g-1}\binom{i}{k-j}$ terms having the form $\left(-1\right)^{g+1}\frac{1}{k-g+1}x_1^{k-j}x_k^j$. The final kth group is the term $\left(-1\right)^{k-1}\left(k-1\right)x_1^i x_k^{k-i}$. So, if $i+j = k$, $j \geq \frac{k}{2}$, $i \leq \frac{k}{2}$, the summed coefficient of $x_1^i x_k^{k-i}$ is $\left(-1\right)^{k-1}\left(k-1\right) + \sum_{g=i+1}^{k-1}\left(-1\right)^{g+1}\frac{1}{k-g+1}\left(k-i\right)\binom{k-i-1}{g-i-1} + \sum_{g=k-i+1}^{k-1}\left(-1\right)^{g+1}\frac{1}{k-g+1}i\binom{i-1}{g-k+i-1} = \left(-1\right)^{k-1}\left(k-1\right) + \left(-1\right)^{k+1} + \left(k-i\right)\left(-1\right)^k + \left(-1\right)^k\left(i-1\right) = \left(-1\right)^{k+1}$. The summation identities are $\sum_{g=i+1}^{k-1}\left(-1\right)^{g+1}\frac{1}{k-g+1}\left(k-i\right)\binom{k-i-1}{g-i-1} = \left(k-i\right)\int_0^1\sum_{g=i+1}^{k-1}\left(-1\right)^{g+1}\binom{k-i-1}{g-i-1}t^{k-g}dt = \left(k-i\right)\int_0^1\left(\left(-1\right)^i\left(t-1\right)^{k-i-1} - \left(-1\right)^{k+1}\right)dt =$

$(k-i)\left(\frac{(-1)^k}{i-k}+(-1)^k\right) = (-1)^{k+1} + (k-i)(-1)^k$ and $\sum_{g=k-i+1}^{k-1}(-1)^{g+1}\frac{1}{k-g+1}i\binom{i-1}{g-k+i-1} = \int_0^1 \sum_{g=k-i+1}^{k-1}(-1)^{g+1} i\binom{i-1}{g-k+i-1}t^{k-g}dt = \int_0^1\left(i(-1)^{k-i}(t-1)^{i-1} - i(-1)^{k+1}\right)dt = (-1)^k(i-1)$. If $j<\frac{k+1-i}{2}$, $i>k-1$, if $i=k$, $\psi_k=0$, if $\frac{k+1-i}{2}\le j\le\frac{k-1}{2}$, $\frac{k+1}{2}\le i\le k-1$, the summed coefficient of $x_1^i x_k^{k-i}$ is $(-1)^{k-1}(k-1)+\sum_{g=k-i+1}^{k-1}(-1)^{g+1}\frac{1}{k-g+1}i\binom{i-1}{g-k+i-1} + \sum_{g=i+1}^{k-1}(-1)^{g+1}\frac{1}{k-g+1}(k-i)\binom{k-i-1}{g-i-1}$, the same as above. If $i+j<k$, since $\binom{i}{k-j}=0$, the related terms can be ignored, so, using the binomial theorem and beta function, the summed coefficient of $x_1^{k-j}x_k^j$ is $\sum_{g=j+1}^{i+j}(-1)^{g+1}\frac{1}{k-g+1}i\binom{i-1}{g-j-1}\binom{k-i}{j} = i\binom{k-i}{j}\int_0^1\sum_{g=j+1}^{i+j}(-1)^{g+1}\binom{i-1}{g-j-1}t^{k-g}dt = \binom{k-i}{j}i\int_0^1\left((-1)^j t^{k-j-1}\left(\frac{t}{t-1}\right)^{1-i}\right)dt = \binom{k-i}{j}i\frac{(-1)^{j+i+1}\Gamma(i)\Gamma(k-j-i+1)}{\Gamma(k-j+1)} = \frac{(-1)^{j+i+1}i!(k-j-i)!(k-i)!}{(k-j)!j!(k-j-i)!} = (-1)^{j+i+1}\frac{i!(k-i)!}{k!}\frac{k!}{(k-j)!j!} = \binom{k}{i}^{-1}(-1)^{1+i}\binom{k}{j}(-1)^j$.

The coefficient of $x_1^i x_k^{k-i}$ in $\binom{k}{i}^{-1}(-1)^{1+i}(x_1-x_k)^k$ is $\binom{k}{i}^{-1}(-1)^{1+i}\binom{k}{i}(-1)^{k-i} = (-1)^{k+1}$, same as the summed coefficient if $i+j=k$. If $i+j<k$, the coefficient of $x_1^{k-j}x_k^j$ is $\binom{k}{i}^{-1}(-1)^{1+i}\binom{k}{j}(-1)^j$, same as the corresponding summed coefficient. Therefore, $\psi_k(x_1=x_1,\ldots,x_i=x_1,x_{i+1}=x_k,\ldots,x_k=x_k) = \binom{k}{i}^{-1}(-1)^{1+i}(x_1-x_k)^k$, the maximum and minimum of ψ_k follow directly from the properties of the binomial coefficient. □

ξ_Δ is closely related to $f_\Xi(\Delta)$, which is the pairwise difference distribution, since the probability density of ξ_Δ can be expressed as $f_{\Xi_k}(\bar{\Delta}|\Delta)$ and $\sum_{\bar{\Delta}=-\binom{k}{\frac{3+(-1)^k}{2}}^{-1}(-\Delta)^k}^{\frac{1}{k}(-\Delta)^k} f_{\Xi_k}(\bar{\Delta}|\Delta) = f_\Xi(\Delta) = \int_0^\infty 2f(t)f(t-\Delta)dt$. The support of the original distribution is assumed to be $[0,\infty)$ for simplicity. Recall that $f_\Xi(\Delta)$ is monotonic increasing with a mode at the origin if the original distribution is unimodal. Thus, in general, ignoring the shape of ξ_Δ, Ξ_k is monotonic left and right around zero. In fact, the median of Ξ_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 Δ 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 Δ 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 Ξ_k will only be slightly shifted from zero. Denote the median of Ξ_k as mkm, for the five parametric distributions here, $|mkm|$s are all $\le 0.1\sigma$ for Ξ_3 and Ξ_4 (SI Dataset S1). Assuming $mkm=0$, 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 $\frac{\frac{1}{2}}{\binom{k}{2}^{-1}(Q(0)-Q(1))^k} > \frac{\frac{1}{2}}{\frac{1}{k}(Q(0)-Q(1))^k}$. This means that, on average, the inequality $f(Q(\epsilon))\ge f(Q(1-\epsilon))$ 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 $\frac{\partial\psi_3(x_1,x_2,x_3)}{\partial x_2} = -\frac{x_1^2}{2} - x_1x_2 + 2x_1x_3 + x_2^2 - x_2x_3 - \frac{x_3^2}{2}$, $-\frac{3}{4}(x_1-x_3)^2 \le \frac{\partial\psi_3(x_1,x_2,x_3)}{\partial x_2} \le -\frac{1}{2}(x_1-x_3)^2 \le 0$. 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 $f(Q(\epsilon))\le f(Q(1-\epsilon))$ holds (the same result can be inferred from the definition of central moments, where the positivity of the odd order central moment is directly related to the left-skewness of the corresponding kernel distribution). In all, the monotonicity of the pairwise difference distribution guides the general shape of the kth central moment kernel distribution, $k>2$, forcing it to be unimodal-like with the mode and median close to zero, then, the inequality $f(Q(\epsilon))\le f(Q(1-\epsilon))$ or $f(Q(\epsilon))\ge f(Q(1-\epsilon))$ holds in general. If a distribution is ordered and all of its central moment kernel distributions are also ordered, it is called completely ordered. Although strict complete orderliness is difficult to prove, even if the inequality may be violated in a small range, as discussed in Subsection A, the mean-SWA-median inequality remains valid, in most cases, for the central moment kernel distribution.

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. $\psi_k(x_1=\lambda x_1+\mu,\cdots,x_k=\lambda x_k+\mu) = \lambda^k\psi_k(x_1,\cdots,x_k)$.

A direct result of Theorem B.3 is that, Wkm after standardization is invariant to location and scale. So, the weighted standardized kth moment is defined to be

$$\mathrm{W}skm_{\epsilon_{U_k},\gamma,n} \coloneqq \frac{\mathrm{W}km_{\epsilon_{U_k},\gamma,n}}{\mathrm{W}var^{\frac{k}{2}}_{\epsilon_{U_2},\gamma,n}}.$$

Consider two continuous distributions belonging to the same location–scale family, their corresponding kth central moment kernel distributions only differ in scaling. So d is invariant, as shown in Subsection A. The recombined kth central moment, based on rm, is defined by,

$$rkm_{d,\epsilon_{U_k},n} \coloneqq (d+1)\,\mathrm{SW}km_{\epsilon_{U_k},n} - dmkm_n,$$

where SW$km_{\epsilon_{U_k},n}$ is using the binomial kth central moment (B$km_{\epsilon_{U_k},n}$) here, mkm_n is the median kth central moment. Since SW$km_{\epsilon_{U_k},n}$ is an L-statistic, the resulting $rkm_{d,\epsilon_{U_k},n}$ is an arithmetic I-statistic. Similarly, the quantile will not change after scaling. The quantile kth central moment is thus defined as

$$qkm_{d,\epsilon_{U_k},n} \coloneqq \hat{Q}_n\left(\left(p\mathrm{SW}km_{\epsilon_{U_k},n}-\frac{1}{2}\right)d + p\mathrm{SW}km_{\epsilon_{U_k},n}\right),$$

where $p\mathrm{SW}km_{\epsilon_{U_k},n} = \hat{F}_{\psi,n}\left(\mathrm{SW}km_{\epsilon_{U_k},n}\right)$, $\hat{F}_{\psi,n}$ is the empirical cumulative distribution function of the corresponding central moment kernel distribution. $qkm_{d,\epsilon_{U_k},n}$ is a quantile I-statistic.

For standardized moments, quantile skewness and quantile kurtosis are defined to be $qskew_{d,\epsilon_{U_3},n} \coloneqq \frac{qtm_{d,\epsilon_{U_3},n}}{qsd^3_{d,\epsilon_{U_2},n}}$ and $qkurt_{d,\epsilon_{U_4},n} \coloneqq \frac{qfm_{d,\epsilon_{U_4},n}}{qsd^4_{d,\epsilon_{U_2},n}}$. Quantile standard deviation ($qsd_{d,\epsilon_{U_2},n}$), recombined standard deviation ($rsd_{d,\epsilon_{U_2},n}$), quantile third central moment ($qtm_{d,\epsilon_{U_3},n}$), quantile fourth central moment ($qfm_{d,\epsilon_{U_4},n}$), recombined third central moment ($rtm_{d,\epsilon_{U_3},n}$), recombined fourth central moment ($rfm_{d,\epsilon_{U_4},n}$), recombined skewness ($rskew_{d,\epsilon_{U_3},n}$), and recombined kurtosis ($rkurt_{d,\epsilon_{U_4},n}$) are all defined similarly as above and not repeated here. The transformation to a location problem can also empower related statistical tests. From the better performance of the quantile mean in heavy-tailed distributions, quantile central moments are generally better than recombined central moments regarding asymptotic bias.

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) (38), 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 (39) is more sophisticated and can overcome the limitation of the number of blocks, but the second kind of bias remains unsolved.

C. Congruent distribution. In the realm of nonparametric statistics, the precise values of robust estimators are of secondary importance. What is of primary importance is their relative differences or orders. Based on this principle, in the absence of contamination, as the parameters of the distribution vary, all reasonable nonparametric location estimates should asymptotically change in the same direction. Otherwise if the results obtained based on the trimmed mean are completely different from those based on the median, a contradiction arises. However, such contradictions are possible, as in the case of the Weibull distribution, $m = \lambda \sqrt[\alpha]{\ln(2)}$, $\mu = \lambda\Gamma\left(1+\frac{1}{\alpha}\right)$, then, when $\alpha = 1$, $m = \lambda\ln(2) \approx 0.693\lambda$, $\mu = \lambda$, but when $\alpha = \frac{1}{2}$, $m = \lambda\ln^2(2) \approx 0.480\lambda$, $\mu = 2\lambda$, the mean increases, but the median decreases. To study the conditions that avoid such scenarios by classifying distributions through the signs of derivatives, let the quantile average function of a parametric distribution be denoted as $\mathrm{QA}\left(\epsilon, \gamma, \alpha_1, \cdots, \alpha_i, \cdots, \alpha_k\right)$, where α_i represent the parameters of the distribution, then, a distribution is γ-congruent if and only if the sign of $\frac{\partial \mathrm{QA}}{\partial \alpha_i}$ remains the same for all $0 \le \epsilon \le \frac{1}{1+\gamma}$. If this partial derivative is equal to zero or undefined, it can be considered both positive and negative, and thus does not impact the analysis. Asymptotically, any weighted average can be expressed as an integral of the quantile average function. Since the sign does not change after integration, the sign of $\frac{\partial \mathrm{QA}}{\partial \alpha_i}$ remains the same for all $0 \le \epsilon \le \frac{1}{1+\gamma}$ implies that all γ-weighted averages change in the same direction as the parameters change, as long as they are not undefined. 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 QHLM gives the definition of γ-U-congruence. Chebyshev's inequality implies that, for any probability distribution with finite moments, even if some weighted averages change in a direction different from that of the sample mean, the deviations are bounded. Furthermore, distributions with infinite moments can be γ-congruent, since the definition is based on the quantile average, not the sample mean.

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

Theorem C.1. *A symmetric distribution with a finite second moment is always congruent.*

Proof. For any symmetric distribution with a finite second moment, all symmetric quantile averages coincide. The conclusion follows immediately. □

Theorem C.2. *A positive define location-scale distribution with a finite second moment is always γ-congruent.*

Proof. As shown in discussions in Subsection A, for a location-scale distribution, any weighted average can be expressed as $\lambda \mathrm{WA}_0(\epsilon) + \mu$, where $\mathrm{WA}_0(\epsilon)$ is an integral of $\mathrm{Q}_0(p)$ according to the definition of the weighted average. 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 with a finite second moment is always γ-congruent.*

Proof. Theorem B.3 shows that the corresponding central moment kernel distribution is also a location-scale family distribution. Theorem B.1 shows that it is positively defined. Implementing Theorem C.2 yields the desired result. □

For the Pareto distribution, $\frac{\partial Q(p,\alpha)}{\partial \alpha} = \frac{x_m(1-p)^{-1/\alpha}\ln(1-p)}{\alpha^2}$. Since $\ln(1-p) < 0$ for all $0 < p < 1$, $(1-p)^{-1/\alpha} > 0$ for all $0 < p < 1$ and $\alpha > 0$, so $\frac{\partial Q(p,\alpha)}{\partial \alpha} < 0$, and therefore $\frac{\partial \mathrm{QA}(\epsilon,\gamma,\alpha)}{\partial \alpha} < 0$, the Pareto distribution is γ-congruent. The derivative for the lognormal distribution is $\frac{\partial \mathrm{SQA}(\epsilon,\sigma)}{\partial \sigma} = \frac{-\mathrm{erfc}^{-1}(2\epsilon)e^{\mu-\sqrt{2}\sigma\mathrm{erfc}^{-1}(2\epsilon)} - \mathrm{erfc}^{-1}(2-2\epsilon)e^{\mu-\sqrt{2}\sigma\mathrm{erfc}^{-1}(2-2\epsilon)}}{\sqrt{2}}$. Since the inverse complementary error function is positive when the input is smaller than 1, and negative when the input is larger than 1, $\mathrm{erfc}^{-1}(2\epsilon) = -\mathrm{erfc}^{-1}(2-2\epsilon)$, $e^{\mu-\sqrt{2}\sigma\mathrm{erfc}^{-1}(2-2\epsilon)} > e^{\mu-\sqrt{2}\sigma\mathrm{erfc}^{-1}(2\epsilon)}$, $\frac{\partial \mathrm{SQA}(\epsilon,\sigma)}{\partial \sigma} > 0$, the lognormal distribution is congruent. Theorem C.1 implies that the generalized Gaussian distribution is 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. The main reason for non-congruence of a right-skewed smooth partial bounded probability distribution lies in the simultaneous increase of probability densities on two opposite sides: one approaching the bound and the other approaching infinity. Note that the gamma distribution does not have this issue, it looks to be congruent.

Although some common parametric distributions are not congruent, Theorem C.2 establishes that γ-congruence always holds for a positive define location-scale family distribution and thus for the second central moment kernel distribution generated from a continuous location-scale unimodal distribution as shown in Theorem C.3. Theorem B.2 demonstrates that all their central moment kernel distributions are unimodal-like with mode and median close to zero, as long as they are unimodal distributions. Assuming finite moments and constant $Q(0) - Q(1)$, increasing the mean of the kernel distribution will result in a more heavy-tailed distribution, i.e., the probability density of the values close to $\frac{1}{k}(-\Delta)^k$ increases. While the total probability density on either side of zero remains unchanged 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. 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. Therefore, they are valid measures for nonparametric descriptive statistics.

D. A shape-scale distribution as the consistent distribution.

Up to this point, in this article, the consistent robust estimation has been limited to a location-scale distribution, with the location parameter often being omitted for simplicity. To construct probability distributions can be made to fit the observed skewness and kurtosis arbitrarily well, in 1894, Pearson (40) introduced a family of continuous probability distributions that are now often characterized by the square of the skewness and the kurtosis. If the skewness and the kurtosis are interrelated by a shape parameter, a distribution specified by a shape parameter (denoted as α) and a scale parameter (denoted as λ) is often referred to as a shape-scale distribution. Weibull, gamma, Pareto, lognormal, and generalized Gaussian distributions (when μ is a constant) are all shape-scale unimodal distributions. Moreover, if α or skewness or kurtosis is a constant, the shape-scale distribution is reduced to a location-scale distribution. The above discussion shows that, due to the invariant property, if a location-scale distribution is chosen as the consistent distribution, the type of invarient moments and their related weighted moments are given, there should exist a unique k-tuple $(d_{im}, \ldots, d_{ikm})$ calibrated by the distribution and the corresponding kernel distributions generated from this distribution. For a right skewed shape-scale distribution, let $D(|skewness|, kurtosis, k, etype, dtype, n) = d_{ikm}$ denote these relations, 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 $k = 1$, 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 due to the invariant property of scale, 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 identical to the skewness for them and it can also handle 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

$$\begin{cases} rm\left(\mathrm{SWL}, m, D(|rskew|, rkurt, 1)\right) = \mu \\ rvar\left(\mathrm{SW}var, mvar, D(|rskew|, rkurt, 2)\right) = \mu_2 \\ rtm\left(\mathrm{SW}tm, mtm, D(|rskew|, rkurt, 3)\right) = \mu_3 \\ rfm\left(\mathrm{SW}fm, mfm, D(|rskew|, rkurt, 4\right) = \mu_4 \\ rskew = \frac{\mu_3}{\mu_2^{\frac{3}{2}}} \\ rkurt = \frac{\mu_4}{\mu_2^2} \end{cases},$$

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 $\varsigma(|rskew|) = \left|\frac{rtm(\mathrm{SW}tm, mtm, D(|rskew|, 3))}{rvar(\mathrm{SW}var, mvar, D(|rskew|, 2))^{\frac{3}{2}}}\right|$ and $\varkappa(rkurt) = \frac{rfm(\mathrm{SW}fm, mfm, D(rkurt, 4))}{rvar(\mathrm{SW}var, mvar, D(rkurt, 2))^2}$. 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; SWvar; SWfm; $mvar$; mfm; $maxit$; δ
Output: $rkurt_{i-1}$
$i = 0$
2: $rkurt_i \leftarrow \varkappa(kurtosis_{max})$ ▷ Using the maximum kurtosis available in D as an initial guess.
repeat
4: $i = i + 1$
$rkurt_{i-1} \leftarrow rkurt_i$
6: $rkurt_i \leftarrow \varkappa(rkurt_{i-1})$
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 validaty of Algorithm 1.

Theorem D.1. *Assuming mkms are all equal to zero, $|rskew|$ and $rkurt$, defined as the largest attracting fix 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 central moment kernel distributions are all 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.*

Proof. Without loss of generality, only $rkurt$ is considered here, while the logic for $|rskew|$ is the same. Also, according to the property of invariance, the second central moments of the underlying distribution of the sample and consistent distribution are all assumed to be 1. From the definition of D, $\frac{\varkappa(rkurt_D)}{rkurt_D} = \frac{\frac{fm_D - \mathrm{SW}fm_D}{\mathrm{SW}fm_D - mfm_D}(\mathrm{SW}fm - mfm) + \mathrm{SW}fm}{rkurt_D\left(\frac{var_D - \mathrm{SW}var_D}{\mathrm{SW}var_D - mvar_D}(\mathrm{SW}var - mvar) + \mathrm{SW}var\right)^2}$, where the subscript D indicates that the estimates are from the central moment kernel distributions generated from the consistent distribution used to calibrate the d values, while other estimates are from the underlying distribution of the sample.

Then, assuming the mkms are all equal to zero, $\frac{\varkappa(rkurt_D)}{rkurt_D} = \frac{\frac{fm_D - \mathrm{SW}fm_D}{\mathrm{SW}fm_D}(\mathrm{SW}fm) + \mathrm{SW}fm}{rkurt_D\left(\frac{\mathrm{SW}var}{\mathrm{SW}var_D}\right)^2} = \frac{\left(\frac{fm_D - \mathrm{SW}fm_D}{\mathrm{SW}fm_D} + 1\right)(\mathrm{SW}fm)}{fm_D\left(\frac{\mathrm{SW}var}{\mathrm{SW}var_D}\right)^2} = \frac{\mathrm{SW}fm\mathrm{SW}var_D^2}{\mathrm{SW}fm_D\mathrm{SW}var^2} = \frac{\frac{\mathrm{SW}fm}{\mathrm{SW}var^2}}{\frac{\mathrm{SW}fm_D}{\mathrm{SW}var_D{}^2}} = \frac{\mathrm{SW}kurt}{\mathrm{SW}kurt_D}$. Since SW$fm_D$ are from the same kernel distribution as $fm_D = rkurt_D var_D{}^2$, according to the congruence, an increase in fm_D will also result in an increase in SWfm_D. Combining with Theorem B.3, SW$kurt$ 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 fix point, let the largest one be fix_{max}, then it is attracting since $\left|\frac{\partial(\varkappa(rkurt_D))}{\partial(rkurt_D)}\right| < 1$ for all $rkurt_D \in [fix_{max}, kurtosis_{max}]$.

Asymptotically, consider any $\mathrm{SW}kurt_D > \mathrm{SW}kurt$, $\frac{\varkappa(rkurt_D)}{rkurt_D} < 1$, the same logic applies, a consistent estimator must be the last attracting fix point, fix_{max} is the consistent estimator. □

As a result of Theorem D.1, assuming continuity, $mkms$ are all equal to zero, and congruence of the central moment kernel distributions, Algorithm 1 converges surely provided that a fix point exists within the domain of D. At this stage, D can only be approximated through a Monte Carlo study. Continuity 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 fix 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 fix point iteration is, in principle, similar to the iterative reweighing in M-estimator, but an advantage of this algorithm is that the optimization is solely related to the d value function and is independent of the sample size (except for the quantile moments, which require re-computation of the quantile function, but this operation has a time complexity of $O(1)$ for a sorted sample). Since $|rskew|$ can specify d_{rm} after optimization, this algorithm enables the robust estimations of all four moments to reach a near-consistent level for common unimodal distributions (Table **??**, 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 (2, 34, 41–48). Daniell (1920) stated (41) that comparing the efficiencies of various kinds of estimators is useless unless they all tend to coincide asymptotically. Bickel and Lehmann, also in the landmark series (47, 48), argued that meaningful comparisons can be made by studying the standardized variances, asymptotic variances, and efficiency bounds of these estimators.

Here, the scaled standard error (SSE) is proposed to estimate the variances of all estimators, including recombined/quantile moments, on a scale more comparable to that of the sample mean.

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 is the main statistic of interest, $\widehat{\theta_m}$, the second to the jth column are $j-1$ 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. 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.

The main statistics of interest here are the sample mean and U-central moment (the central moment estimated by using U-statistics), which is essentially the mean of the central moment kernel distribution, so its standard error should be generally close to $\frac{\sigma_{km}}{\sqrt{n}}$, where σ_{km} is the asymptotic standard deviation of the kernel distribution. Noted that, 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 when the mean value is close to zero, the scaled standard error will approach infinity and therefore be too sensitive to small changes.

The SSEs of all robust estimators proposed here are often, although many exceptions exist, between those of the sample median and median central moments and those of the sample mean and U-central moments (SI Dataset S1). This is because similar monotonic relations between robustness and variance are also very common, e.g., Bickel and Lehmann (48) 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) (49, 50) 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 (51). Another approach can be dated back to Laplace (1812) (52) is using $w\bar{x} + (1-w)m_n$ as a location estimator and w is deduced to achieve optimal variance; examples for symmetric distributions see Samuel-Cahn (1994), Chan and He (1994), and Damilano and Puig (2004)'s papers (53–55). In this study, for robust mean estimation, 22 possible combinations were created using two type of invariant means and related symmetric weighted L-statistics ($\mathrm{WHLM}_{k=5,\epsilon=\frac{1}{8},n}$, $\mathrm{THLM}_{k=5,\epsilon=\frac{1}{8},n}$, $\mathrm{WHLM}_{k=2,\epsilon=\frac{1}{8},n}$, $\mathrm{THLM}_{k=2,\epsilon=\frac{1}{8},n}$, H-L, $\mathrm{BM}_{\epsilon=\frac{1}{8},n}$, $\mathrm{SQM}_{\epsilon=\frac{1}{8},n}$, $\mathrm{BM}_{\nu=2,\epsilon=\frac{1}{8},n}$, $\mathrm{WM}_{\epsilon=\frac{1}{8},n}$, $\mathrm{BWM}_{\epsilon=\frac{1}{8},n}$, and $\mathrm{TM}_{\epsilon=\frac{1}{8},n}$ used here). Each combination has a SSE for a single-parameter distribution, which can be inferred through a Monte Carlo study. Then, the combination with the smallest SSE is chosen (if the percentiles of quantile moments exceed the breakdown point, this combination will be excluded). Similar to Subsection D, let $I(|skewness|, kurtosis, k, dtype, n) = ikm_{\mathrm{WA}}$ denote these relations for all invariant moments. Then, since $\lim_{rkurt \to \infty} \frac{I(rkurt,4)}{I(rkurt,2)^2 rkurt} < 1$, the same fix point iteration algorithm can be used to choose the variance-optimum combinations. The only difference is that unlike D, I is defined to be discontinuous but linear interpolation can also ensure continuity. This approach yields results that are often nearly optimal (SI Dateset S1).

Due to combinatorial explosion, the bootstrap (56), introduced by Efron in 1979, is indispensable for computing invariant central moments in practice. In 1981, Bickel and Freedman (57) 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) (58). In the previous article, the advantages of quasi-bootstrap were discussed (59–61). By using quasi-sampling, the impact of the number of repetitions of the bootstrap, or bootstrap size, on variance is negligible (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 but this can be corrected by recalibrating the d values (SI Text). The default bootstrap size is set as 18 thousand here, as it balances computational cost and finite sample bias, except for the asymptotic value calculation. In general, the variances of invariant central moments are much smaller than those of corresponding unbiased sample central moments (deduced by Cramér (62)), except that of the corresponding second central moment (Table **??**).

F. Robustness. The measure of robustness to gross errors used in this paper is the breakdown point proposed by Hampel (63) in 1968. Previous work has shown that the median of means (MoM) is asymptotically equivalent to the median Hodge-Lehmann mean. Therefore it is also biased for any asymmetric distribution. Nevertheless, the concentration bound of MoM depends on $\sqrt{\frac{1}{n}}$ (64), so 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). An estimator $\hat{\theta}$ has a sample-dependent breakdown point if and only if its asymptotic breakdown point $\epsilon(\hat{\theta}, R, \zeta, \upsilon)$ is zero and the empirical influence function of $\hat{\theta}$ is bounded, where R is the measure of badness, ζ is the contaminating processes, υ is the uncontaminated process. For a full formal definition of the asymptotic breakdown point, which is the breakdown point when $n \to \infty$, and the empirical influence function, the reader is referred to Genton and Lucas (2003) and Devlin, Gnanadesikan and Kettenring (1975)'s papers (65, 66).

Bear in mind that it differs from the "infinitesimal robustness" defined by Hampel, which is related to whether the asymptotic influence function is bounded (67–69). 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.

Furthermore, for weighted L-statistics, separating the breakdown point into upper and lower parts is necessary.

Definition F.2 (Upper/lower breakdown point). The upper breakdown point is the breakdown point generalized in Davies and Gather (2005)'s paper (70). The finite-sample upper breakdown point is the finite sample breakdown point defined by Donoho and Huber (1983) (71) and also detailed in (70). The (finite-sample) lower breakdown point is replacing the infinity symbol in these definitions with negative infinity.

For the robust estimations of central moments or other weighted U-statistics based on a robust location estimator, 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 (71).

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

Proof. Suppose m arbitrary large contaminants are added to the sample. The fraction of bad values in the sample can be represented as $\epsilon_{U_k} = \frac{m}{n+m}$, where n denotes the original number of data points that remain unaffected. In the kernel distribution, $\binom{n}{k}$ out of a total of $\binom{n+m}{k}$ points are not corrupted. Then, the breakdown can be avoided if the following inequality holds

$$\binom{n}{k} > \left(\frac{1}{\epsilon}-1\right) \times \left(\binom{n+m}{k} - \binom{n}{k}\right).$$

Since ϵ is the upper breakdown point of the corresponding weighted L-statistic, $\frac{1}{1+\gamma} \geq \epsilon \geq 0$,

$$\frac{1}{1-\epsilon} > \frac{\binom{n+m}{k}}{\binom{n}{k}} = \frac{(n+m)(n+m-1)\dots(n+m-k+1)}{n(n-1)\dots(n-k+1)}.$$

Assuming $n \to \infty$, k is a constant, $\lim_{n\to\infty}\left(\frac{n+m-k+1}{n-k+1}\right) = \frac{n+m}{n}$, then the above inequality does not hold when $\frac{n+m}{n} \geq \left(\frac{1}{1-\epsilon}\right)^{\frac{1}{k}}$. So, the upper asymptotic breakdown point of the weighted U-statistic is $\epsilon_{U_k} = \frac{m}{n+m} = 1 - \frac{n}{n+m} = 1-(1-\epsilon)^{\frac{1}{k}}$. ☐

Remark. If $k=1$, $1-(1-\epsilon)^{\frac{1}{k}} = \epsilon$, so this formula also holds for the weighted L-statistic itself. When $\epsilon = \frac{1}{2}$, $\gamma = 1$, the weighted U-statistic becomes U-quantile, or median U-statistic, which converges almost surely as proven by Choudhury and Serfling (37) in 1988. Here, to ensure the breakdown points of all four moments are the same, $\frac{1}{16}$, since $\epsilon = 1-(1-\epsilon_{U_k})^k$, the breakdown points of all symmetric weighted L-statistics for the second, third, and fourth central moment estimations are adjusted as $\epsilon = \frac{31}{256}, \frac{721}{4096}, \frac{14911}{65536}$, 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 under regularity conditions, i.e., although all estimators can be biased under departures from the assumptions, but their standardized maximum biases can differ substantially (15, 64, 72–75). Previously, it is shown that another way to qualitatively compare the estimators' robustness to departures from the symmetry assumption is constructing and comparing corresponding semiparametric models. An estimator based on a smaller model is naturally more robust to distributional shift within that model. While this comparison is limited to the smaller semiparametric model and is not universal, it is still valid for a wide range of parametric distributions. Bias bounds are more universal since they can be deduced for distributions with finite moments without assumming unimodality (72, 73). However, bias bounds are often hard to deduce for complex estimators. Also, sometimes there are discrepencies between maximum bias and average bias. For example, the maximum bias of $rm_{d\approx 0.000,\epsilon=\frac{1}{16}}$ is greater than that of $\mathrm{SQM}_{\frac{1}{16}}$ in the gamma distribution, but it has much smaller average biases (SI Dataset S1). 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 asymptotic bias is thus defined as follows.

Definition F.3 (Average asymptotic bias). For a single-parameter distribution, the average asymptotic bias (AAB) is just the asymptotic bias $\frac{\left|\hat{\theta}-\theta\right|}{\sigma}$, where $\hat{\theta}$ is the estimation of θ, σ is the standard deviation of the distribution, if $\hat{\theta}$ is a location estimator, or the standard deviation of the kernel distribution (σ_{km}) , if $\hat{\theta}$ is a central moment estimator. For a two-parameter distribution, the first step is setting the lower bound of the kurtosis range of interest $\tilde{\mu}_{4_l}$. Then, the average asymptotic bias is defined as

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

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 under different distributions. The estimation of central moments based on the location estimations of the kernel distributions also enables the standardization of biases of weighted central moments. Currently, there are several options available, such as using the root mean square deviation from the mode (as in Gauss (1)) or the mean absolute deviation, but the standard deviation is preferred because of its central role in standard error estimation.

In Table **??**, $\delta = 0.1$, $C = 120$. 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 SWLs and SWkms 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 in the AAB computations 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 $\mathrm{TM}_{\frac{1}{16}}$ and $\mathrm{WM}_{\frac{1}{16}}$, 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 (72, 73).

Discussion

Moments, including raw moments, central moments, and standardized moments, are fundamental parameters that determine 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 (76). The strong law of large numbers (proven by Kolmogorov in 1933) (77) 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) (78), Pillai and Meng (2016) (79), Cohen, Davis, and Samorodnitsky (2020) (80), and Brown, Cohen, Tang, and Yam (2021) (81). Lindquist and Rachev (2021) raised a critical question: "What are the proper measures for the location, spread, asymmetry, and dependence (association) for random samples with infinite mean?" (82) in their inspiring comment to Brown et al's paper (81). They suggested using median, interquartile range, and medcouple (83) as the robust versions of the first three standardized moments (84–86). This is not the focus of this paper, but it is almost sure that the estimators proposed here will have a place. Since the efficiency of an L-statistic to the sample mean is generally monotonic with respect to the breakdown point (48), and the estimation of central moments can be transformed into a location estimation problem, similar monotonic relations can be expected. For distributions with infinite moments, 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 most robust version of L-moment (87) being trimmed L-moment (88), moments now also have their standard most robust version based on the complete congruence of the underlying distribution.

More generally, statistics, the theory of analyzing data through the use of probability models and measures of random variables, has evolved over time, with various approaches emerging to meet challenges in practice. While the early development of statistics was focused on parametric methods, there were two main approaches to point estimation. The Gauss–Markov theorem (1, 89) states the principle of minimum variance unbiased estimation which was further enriched by Neyman (1934) (90), Rao (1945) (91), Blackwell (1947) (92), Lehmann and Scheffé (1950, 1955) (49, 50). Maximum likelihood was first introduced by Fisher in 1922 (93) in a multinomial model and later generalized by Cramér (1946), Hájek (1970), and Le Cam (1972) (46, 62, 94). In 1939, Wald (95) combined these two principles and suggested the use of minimax estimates. Hodges and Lehmann in 1950 (96) expanded upon this concept and obtained minimax estimates for a series of important problems. It was soon clear that a minimax estimator should be a Bayes estimator with regard to the least favorable prior distribution of θ as a minimax estimator is the best in the worst case scenario. Following Huber's seminal work (4), M-statistics have dominated the field of parametric robust statistics for over half a century. Non-parametric methods, e.g., the Kolmogorov–Smirnov test, Mann-Whitney-Wilcoxon Test, and Hoeffding's independence test, emerged as a popular alternative to parametric methods in 1950s, as they do not make specific assumptions about the underlying distribution of the data. Hodges and Lehmann in 1956, (97) investigated the asymptotic power properties of the Wilcoxon test (98) using Pitman efficiency (99) and surprisingly revealed that its efficiency bound is nearly optimal. In 1963, they proposed a class of efficient nonparametric location estimators based on the confidence bounds of rank tests (100). In particular, Bickel showed in 1965 that the H-L estimator had nearly optimal efficiency (101). In the previous paper, when compared to other semiparametric mean estimators with the same breakdown point, the H-L estimator was shown to be the bias-optimal choice. This supports with Bickel's conclusion that, in any situations where the degree of contamination and type of distribution are unknown, the H-L estimator is optimal (101). The formal study of semiparametric models was initiated by Stein (102) in 1956. Bickel, in 1982, simplified the general heuristic necessary condition proposed by Stein (102) and derived sufficient conditions for this type of problem, adaptive estimation (103). It has become increasingly apparent that many previously called "nonparametric" models are essentially semiparametric models, 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 (104). As suggested by the title *Parametric Robustness: Small Biases can be Worthwhile*, he showed biases exists, but by carefully designing the estimators, they 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 seemingly "wrong" parametric model belongs to the large semiparametric model but not to the semiparametric/parametric models used for calibration, their performance might still be near-optimal due to the common nature shared by the models used by the estimators. The models can be expanded directly and are not limited to a single parametric form. As such, invariant moments may hold the key to unlocking a new branch of statistical inference that is reliable, flexible, and robust.

Data Availability. Data for Table **??** are given in SI Dataset S1. All codes have been deposited in GitHub.

ACKNOWLEDGMENTS. I gratefully acknowledge the constructive comments made by the editor which substantially improved the clarity and quality of this paper.

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. S Newcomb, Researches on the motion of the moon. part ii, the mean motion of the moon and other astronomical elements derived from observations of eclipses and occultations extending from the period of the babylonians until ad 1908. *United States. Naut. Alm. Off. Astron. paper; v. 9* **9**, 1 (1912).
4. PJ Huber, Robust estimation of a location parameter. *Ann. Math. Stat.* **35**, 73–101 (1964).
5. X He, WK Fung, Method of medians for lifetime data with weibull models. *Stat. medicine* **18**, 1993–2009 (1999).
6. M Menon, Estimation of the shape and scale parameters of the weibull distribution. *Technometrics* **5**, 175–182 (1963).
7. SD Dubey, Some percentile estimators for weibull parameters. *Technometrics* **9**, 119–129 (1967).
8. KM Hassanein, Percentile estimators for the parameters of the weibull distribution. *Biometrika* **58**, 673–676 (1971).
9. NB Marks, Estimation of weibull parameters from common percentiles. *J. applied Stat.* **32**, 17–24 (2005).
10. K Boudt, D Caliskan, C Croux, Robust explicit estimators of weibull parameters. *Metrika* **73**, 187–209 (2011).
11. SD Dubey, *Contributions to statistical theory of life testing and reliability.* (Michigan State University of Agriculture and Applied Science. Department of statistics), (1960).
12. LJ Bain, CE Antle, Estimation of parameters in the weibdl distribution. *Technometrics* **9**, 621–627 (1967).
13. RV Hogg, Adaptive robust procedures: A partial review and some suggestions for future applications and theory. *J. Am. Stat. Assoc.* **69**, 909–923 (1974).
14. RJ Hyndman, Y Fan, Sample quantiles in statistical packages. *The Am. Stat.* **50**, 361–365 (1996).
15. C Bernard, R Kazzi, S Vanduffel, Range value-at-risk bounds for unimodal distributions under partial information. *Insur. Math. Econ.* **94**, 9–24 (2020).
16. WR van Zwet, *Convex Transformations of Random Variables: Nebst Stellingen.* (1964).
17. AL Bowley, *Elements of statistics.* (King) No. 8, (1926).
18. RA Groeneveld, G Meeden, Measuring skewness and kurtosis. *J. Royal Stat. Soc. Ser. D (The Stat.* **33**, 391–399 (1984).
19. PJ Bickel, EL Lehmann, Descriptive statistics for nonparametric models iv. spread in *Selected Works of EL Lehmann.* (Springer), pp. 519–526 (2012).
20. PJ Rousseeuw, C Croux, Alternatives to the median absolute deviation. *J. Am. Stat. association* **88**, 1273–1283 (1993).
21. PJ Bickel, EL Lehmann, Descriptive statistics for nonparametric models. iii. dispersion in *Selected works of EL Lehmann.* (Springer), pp. 499–518 (2012).
22. P Janssen, R Serfling, N Veraverbeke, Asymptotic normality of u-statistics based on trimmed samples. *J. statistical planning inference* **16**, 63–74 (1987).
23. J Hodges, E Lehmann, Matching in paired comparisons. *The Annals Math. Stat.* **25**, 787–791 (1954).
24. S Dharmadhikari, K Jogdeo, Unimodal laws and related in *A Festschrift For Erich L. Lehmann.* (CRC Press), p. 131 (1982).
25. AY Khintchine, On unimodal distributions. *Izv. Nauchno-Isled. Inst. Mat. Mech.* **2**, 1–7 (1938).
26. S Purkayastha, Simple proofs of two results on convolutions of unimodal distributions. *Stat. & probability letters* **39**, 97–100 (1998).
27. RA Fisher, Moments and product moments of sampling distributions. *Proc. Lond. Math. Soc.* **2**, 199–238 (1930).
28. PR Halmos, The theory of unbiased estimation. *The Annals Math. Stat.* **17**, 34–43 (1946).
29. W Hoeffding, A class of statistics with asymptotically normal distribution. *The Annals Math. Stat.* **19**, 293–325 (1948).
30. PM Heffernan, Unbiased estimation of central moments by using u-statistics. *J. Royal Stat. Soc. Ser. B (Statistical Methodol.* **59**, 861–863 (1997).
31. D Fraser, Completeness of order statistics. *Can. J. Math.* **6**, 42–45 (1954).
32. AJ Lee, *U-statistics: Theory and Practice.* (Routledge), (2019).
33. RJ Serfling, Generalized l-, m-, and r-statistics. *The Annals Stat.* **12**, 76–86 (1984).
34. P Janssen, R Serfling, N Veraverbeke, Asymptotic normality for a general class of statistical functions and applications to measures of spread. *The Annals Stat.* **12**, 1369–1379 (1984).
35. MG Akritas, Empirical processes associated with v-statistics and a class of estimators under random censoring. *The Annals Stat.* **14**, 619–637 (1986).
36. I Gijbels, P Janssen, N Veraverbeke, Weak and strong representations for trimmed u-statistics. *Probab. theory related fields* **77**, 179–194 (1988).
37. J Choudhury, R Serfling, Generalized order statistics, bahadur representations, and sequential nonparametric fixed-width confidence intervals. *J. Stat. Plan. Inference* **19**, 269–282 (1988).
38. E Joly, G Lugosi, Robust estimation of u-statistics. *Stoch. Process. their Appl.* **126**, 3760–3773 (2016).
39. P Laforgue, S Clémençon, P Bertail, On medians of (randomized) pairwise means in *International Conference on Machine Learning.* (PMLR), pp. 1272–1281 (2019).
40. K Pearson, Contributions to the mathematical theory of evolution. *Philos. Transactions Royal Soc. London. A* **185**, 71–110 (1894).
41. P Daniell, Observations weighted according to order. *Am. J. Math.* **42**, 222–236 (1920).
42. F Mosteller, On some useful" inefficient" statistics. *The Annals Math. Stat.* **17**, 377–408 (1946).
43. CR Rao, *Advanced statistical methods in biometric research.* (Wiley), (1952).
44. PJ Bickel, , et al., Some contributions to the theory of order statistics in *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability.* Vol. 1, pp. 575–591 (1967).
45. H Chernoff, JL Gastwirth, MV Johns, Asymptotic distribution of linear combinations of functions of order statistics with applications to estimation. *The Annals Math. Stat.* **38**, 52–72 (1967).
46. L LeCam, On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *The Annals Math. Stat.* **41**, 802–828 (1970).
47. P Bickel, E Lehmann, Descriptive statistics for nonparametric models i. introduction in *Selected Works of EL Lehmann.* (Springer), pp. 465–471 (2012).
48. PJ Bickel, EL Lehmann, Descriptive statistics for nonparametric models ii. location in *selected works of EL Lehmann.* (Springer), pp. 473–497 (2012).

49. EL Lehmann, H Scheffé, Completeness, similar regions, and unbiased estimation-part i in *Selected works of EL Lehmann.* (Springer), pp. 233–268 (2011).
50. EL Lehmann, H Scheffé, *Completeness, similar regions, and unbiased estimation—part II.* (Springer), (2012).
51. 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).
52. PS Laplace, *Theorie analytique des probabilities.* (1812).
53. E Samuel-Cahn, Combining unbiased estimators. *The Am. Stat.* **48**, 34 (1994).
54. Y Chan, X He, A simple and competitive estimator of location. *Stat. & Probab. Lett.* **19**, 137–142 (1994).
55. G Damilano, P Puig, Efficiency of a linear combination of the median and the sample mean: The double truncated normal distribution. *Scand. J. Stat.* **31**, 629–637 (2004).
56. B Efron, Bootstrap methods: Another look at the jackknife. *The Annals Stat.* **7**, 1–26 (1979).
57. PJ Bickel, DA Freedman, Some asymptotic theory for the bootstrap. *The annals statistics* **9**, 1196–1217 (1981).
58. R Helmers, P Janssen, N Veraverbeke, *Bootstrapping U-quantiles.* (CWI. Department of Operations Research, Statistics, and System Theory [BS]), (1990).
59. 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).
60. 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).
61. KA Do, P Hall, Quasi-random resampling for the bootstrap. *Stat. Comput.* **1**, 13–22 (1991).
62. H Cramér, *Mathematical methods of statistics.* (Princeton university press) Vol. 43, (1999).
63. FR Hampel, *Contributions to the theory of robust estimation.* (University of California, Berkeley), (1968).
64. L Devroye, M Lerasle, G Lugosi, RI Oliveira, Sub-gaussian mean estimators. *The Annals Stat.* **44**, 2695–2725 (2016).
65. MG Genton, A Lucas, Comprehensive definitions of breakdown points for independent and dependent observations. *J. Royal Stat. Soc. Ser. B (Statistical Methodol.* **65**, 81–94 (2003).
66. SJ Devlin, R Gnanadesikan, JR Kettenring, Robust estimation and outlier detection with correlation coefficients. *Biometrika* **62**, 531–545 (1975).
67. FR Hampel, A general qualitative definition of robustness. *The annals mathematical statistics* **42**, 1887–1896 (1971).
68. FR Hampel, The influence curve and its role in robust estimation. *J. american statistical association* **69**, 383–393 (1974).
69. PJ Rousseeuw, FR Hampel, EM Ronchetti, WA Stahel, *Robust statistics: the approach based on influence functions.* (John Wiley & Sons), (2011).
70. PL Davies, U Gather, Breakdown and groups. *The Annals Stat.* **33**, 977 – 1035 (2005).
71. DL Donoho, PJ Huber, The notion of breakdown point. *A festschrift for Erich L. Lehmann* **157184** (1983).
72. M Bieniek, Comparison of the bias of trimmed and winsorized means. *Commun. Stat. Methods* **45**, 6641–6650 (2016).
73. K Danielak, T Rychlik, Theory & methods: Exact bounds for the bias of trimmed means. *Aust. & New Zealand J. Stat.* **45**, 83–96 (2003).
74. S Minsker, Uniform bounds for robust mean estimators. *arXiv preprint arXiv:1812.03523* (2018).
75. T Mathieu, Concentration study of m-estimators using the influence function. *Electron. J. Stat.* **16**, 3695–3750 (2022).
76. PL Hsu, H Robbins, Complete convergence and the law of large numbers. *Proc. national academy sciences* **33**, 25–31 (1947).
77. A Kolmogorov, Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.* **4**, 83–91 (1933).
78. M Drton, H Xiao, Wald tests of singular hypotheses. *Bernoulli* **22**, 38–59 (2016).
79. NS Pillai, XL Meng, An unexpected encounter with cauchy and lévy. *The Annals Stat.* **44**, 2089–2097 (2016).
80. 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).
81. 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).
82. WB Lindquist, ST Rachev, Taylor's law and heavy-tailed distributions. *Proc. Natl. Acad. Sci.* **118**, e2118893118 (2021).
83. G Brys, M Hubert, A Struyf, A robust measure of skewness. *J. Comput. Graph. Stat.* **13**, 996–1017 (2004).
84. DC Hoaglin, F Mosteller, JW Tukey, *Exploring data tables, trends, and shapes.* (John Wiley & Sons), (2011).
85. PJ Huber, *Robust statistics.* (Wiley), pp. 309–312 (1981).
86. RA Maronna, RD Martin, VJ Yohai, M Salibián-Barrera, *Robust statistics: theory and methods (with R).* (John Wiley & Sons), (2019).
87. 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).
88. EA Elamir, AH Seheult, Trimmed l-moments. *Comput. Stat. & Data Analysis* **43**, 299–314 (2003).
89. AA Markov, *Wahrscheinlichkeitsrechnung.* (Teubner), (1912).
90. 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).
91. C Radhakrishna Rao, Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **37**, 81–91 (1945).
92. D Blackwell, Conditional expectation and unbiased sequential estimation. *The Annals Math. Stat.* pp. 105–110 (1947).
93. 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).
94. 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).
95. A Wald, Contributions to the theory of statistical estimation and testing hypotheses. *The Annals Math. Stat.* **10**, 299–326 (1939).
96. J Hodges, EL Lehmann, Some problems in minimax point estimation in *Selected Works of EL Lehmann.* (Springer), pp. 15–30 (2012).
97. JL Hodges Jr, EL Lehmann, The efficiency of some nonparametric competitors of the t-test. *The Annals Math. Stat.* pp. 324–335 (1956).
98. F Wilcoxon, Individual comparisons by ranking methods. *Biom. Bull.* **1**, 80–83 (1945).
99. EJ Pitman, *Lecture Notes on Nonparametric Statistical Inference: Lectures Given for the University of North Carolina,[Chapel Hill], 1948.* (University of North Carolina), (1948).
100. J Hodges Jr, E Lehmann, Estimates of location based on rank tests. *The Annals Math. Stat.* **34**, 598–611 (1963).
101. PJ Bickel, On some robust estimates of location. *The Annals Math. Stat.* **36**, 847–858 (1965).
102. C Stein, , et al., Efficient nonparametric testing and estimation in *Proceedings of the third Berkeley symposium on mathematical statistics and probability.* Vol. 1, pp. 187–195 (1956).
103. PJ Bickel, On adaptive estimation. *The Annals Stat.* **10**, 647–671 (1982).
104. P Bickel, Parametric robustness: small biases can be worthwhile. *The Annals Stat.* **12**, 864–879 (1984).