

Near-consistent robust estimations of moments for unimodal distributions

Tuban Lee^{a,1}

^aInstitute of Biomathematics, Macau SAR 999078, China

This manuscript was compiled on April 1, 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 $I(LE_1, \dots, LE_l) = (\theta_1, \dots, \theta_q)$ 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 θ s 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 (BHL M_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 ??) 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.

¹To 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

$$rm_{d,\epsilon,n} := \lim_{c \rightarrow \infty} \left(\frac{(\text{SWL}_{\epsilon,n} + c)^{d+1}}{(m_n + c)^d} - c \right),$$

where d is the key factor for bias correction, m_n is the sample median, $\text{SWL}_{\epsilon,n}$ is $\text{BM}_{\epsilon,n}$ 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 $\alpha \rightarrow \infty$.*

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, $rm_{d,\epsilon} = \lim_{c \rightarrow \infty} \left(\frac{(\text{BM}_{\epsilon,n} + c)^{d+1}}{(m_n + c)^d} - c \right) = (d+1)\text{BM}_{\epsilon} - dm = \mu$. So, $d = \frac{\mu - \text{BM}_{\epsilon}}{\text{BM}_{\epsilon} - m}$. 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, $E\left[\text{BM}_{\frac{1}{8},n}\right] = \lambda \left(1 + \ln\left(\frac{46656}{8575\sqrt{35}}\right)\right)$. 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, $E[\text{BM}_{\epsilon,n}] = E[m_n] = E[X]$. Then $E[rm_{d,\epsilon,n}] = \lim_{c \rightarrow \infty} \left(\frac{(E[X] + c)^{d+1}}{(E[X] + c)^d} - c \right) = E[X]$. 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 $d_{\text{Pareto}} = \frac{\frac{\alpha x_m}{\alpha - 1} - x_m(1-p_1)^{-\frac{1}{\alpha}}}{x_m(1-p_1)^{-\frac{1}{\alpha}} - x_m(1-p_2)^{-\frac{1}{\alpha}}}$. Since any weighted L -statistic can be expressed as an integral of the quantile function, $\lim_{\alpha \rightarrow \infty} \frac{\frac{\alpha}{1-p_1} - (1-p_1)^{-1/\alpha}}{(1-p_1)^{-1/\alpha} - (1-p_2)^{-1/\alpha}} = -\frac{\ln(1-p_1)+1}{\ln(1-p_1)-\ln(1-p_2)}$, the d value for the Pareto distribution approaches that of the exponential distribution as $\alpha \rightarrow \infty$, regardless of the type of weighted L -statistic used. This completes the demonstration. \square

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, $F(x) = Q^{-1}(x) \rightarrow x = Q(p) = \lambda Q_0(p) + \mu$. Thus, any L -statistic can be expressed as $\lambda L_0(\epsilon) + \mu$, where $L_0(\epsilon)$ is an integral of $Q_0(p)$ according to the definition of the L -statistic. The simultaneous cancellation of μ and λ in $\frac{(\lambda\mu_0 + \mu) - (\lambda L_0(\epsilon) + \mu)}{(\lambda L_0(\epsilon) + \mu) - (\lambda m_0 + \mu)}$ 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

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

provided that $\hat{F}_n(\text{SWL}_{\epsilon,n}) \geq \frac{1}{2}$, 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, $\hat{F}_n(x) := \frac{1}{n} \left(\frac{x - X_{sp}}{X_{sp+1} - X_{sp}} + sp \right)$, where $sp = \sum_{i=1}^n \mathbf{1}_{X_i \leq x}$, $\mathbf{1}_A$ is the indicator of event A . The solution of $\hat{F}_n(\text{SWL}_{\epsilon,n}) < \frac{1}{2}$ is reversing the percentile by $1 - \hat{F}_n(\text{SWL}_{\epsilon,n})$, the obtained percentile is also reversed. Without loss of generality, in the following discussion, only the case where $\hat{F}_n(\text{SWL}_{\epsilon,n}) \geq \frac{1}{2}$ 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 ($\alpha \rightarrow \infty$) and any symmetric distributions provided that the second moments are finite.*

Proof. Similarly, rearranging the definition, $d = \frac{F(\mu) - F(\text{BM}_{\epsilon})}{F(\text{BM}_{\epsilon}) - \frac{1}{2}}$. The cdf of the exponential distribution is $F(x) = 1 - e^{-\lambda^{-1}x}$, $\lambda \geq 0$, $x \geq 0$, the expectation of $\text{BM}_{\epsilon,n}$ can be expressed as $\lambda \text{BM}_0(\epsilon)$, so $F(\text{BM}_{\epsilon})$ 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, $F(E[\text{BM}_{\epsilon,n}]) = F(\mu) = \frac{1}{2}$. Then, the expectation of the quantile mean is $qm_{d,\epsilon} = F^{-1}\left(\left(F(\mu) - \frac{1}{2}\right)d + F(\mu)\right) = F^{-1}\left(0 + F(\mu)\right) = \mu$.

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

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

153 $\frac{1 - (\frac{\alpha-1}{\alpha})^{\alpha-p_1}}{p_1-p_2}$. When $\alpha \rightarrow \infty$, $(\frac{\alpha-1}{\alpha})^{\alpha} = \frac{1}{e}$. The d value
 154 for the exponential distribution is identical, since $d_{exp} =$

$$\frac{(1-e^{-1}) - \left(1 - e^{-\ln\left(\frac{1}{1-p_1}\right)}\right)}{\left(1 - e^{-\ln\left(\frac{1}{1-p_1}\right)}\right) - \left(1 - e^{-\ln\left(\frac{1}{1-p_2}\right)}\right)} = \frac{1 - \frac{1}{e} - p_1}{p_1 - p_2}$$

 155 All results
 156 are now proven. \square

157 The definitions of location and scale parameters are such
 158 that they must satisfy $F(x; \lambda, \mu) = F(\frac{x-\mu}{\lambda}; 1, 0)$. By recalling
 159 $x = \lambda Q_0(p) + \mu$, it follows that the percentile of any weighted
 160 L -statistic is free of λ and μ , which guarantees the validity of
 161 the quantile mean. The quantile mean is a quantile I -statistic.
 162 Specifically, an estimator is classified as a quantile I -statistic
 163 if LEs are percentiles of a distribution obtained by plugging
 164 L -statistics into a cumulative distribution function and I is
 165 defined with arithmetic operations, constants and quantile
 166 functions. $qm_{d \approx 0.321, \epsilon = \frac{1}{8}}$ works better in the fat-tail scenarios
 167 (SI Dataset S1). Theorem .1 and A.1 show that $rm_{d \approx 0.375, \epsilon = \frac{1}{8}}$
 168 and $qm_{d \approx 0.321, \epsilon = \frac{1}{8}}$ are both consistent mean estimators for
 169 any symmetric distribution and a skewed distribution with
 170 finite second moments. It's obvious that the breakdown points
 171 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
 172 they are all invariant means.

173 To study the impact of the choice of SWAs in rm and
 174 qm , it is constructive to recall that a symmetric weighted
 175 average is a linear combination of symmetric quantile aver-
 176 ages. While using a less-biased symmetric weighted average
 177 can generally enhance performance (SI Dataset S1), there is
 178 a greater risk of violation in the semiparametric framework.
 179 However, the mean-SWA $_{\epsilon}$ -median inequality is robust to slight
 180 fluctuations of the SQA function of the underlying distribu-
 181 tion. Suppose the SQA function is generally decreasing in
 182 $[0, u]$, but increasing in $[u, \frac{1}{2}]$, since $1 - 2\epsilon$ of the symmet-
 183 ric quantile averages will be included in the computation of
 184 SWA $_{\epsilon}$, as long as $\frac{1}{2} - u \ll 1 - 2\epsilon$, and other portions of the
 185 SQA function satisfy the inequality constraints that define
 186 the ν th orderliness on which the SWA $_{\epsilon}$ is based, the mean-
 187 SWA $_{\epsilon}$ -median inequality will still hold. This is due to the
 188 violation being bounded (15) and therefore cannot be extreme
 189 for unimodal distributions. For instance, the SQA function is
 190 non-monotonic when the shape parameter of the Weibull dis-
 191 tribution $\alpha > \frac{1}{1-\ln(2)} \approx 3.259$ as shown in the previous article,
 192 the violation of the third orderliness starts near this parameter
 193 as well, yet the mean-BM $_{\frac{1}{8}}$ -median inequality is still valid
 194 when $\alpha \leq 3.322$. Another key factor in determining the risk
 195 of violation is the skewness of the distribution. Previously, it
 196 was demonstrated that in a family of distributions differing
 197 by a skewness-increasing transformation in van Zwet's sense,
 198 the violation of orderliness, if it happens, often only occurs
 199 when the distribution is nearly symmetrical (16). The over-
 200 corrections in rm and qm are dependent on the SWA $_{\epsilon}$ -median
 201 difference, which can be a reasonable measure of skewness
 202 (17, 18), implying that the over-correction is often tiny with
 203 a moderate d . The sample logic can be applied to SWHLM.
 204 This qualitative analysis provides another perspective, in ad-
 205 dition to the bias bounds (15), that rm and qm based on the
 206 mean-SWL $_{\epsilon}$ -median inequality are generally safe.

207 **B. Robust estimations of the central moments.** In 1979, Bickel
 208 and Lehmann, in their final paper of the landmark series *De-*

209 *scriptive Statistics for Nonparametric Models* (19), generalized
 210 a class of estimators called "measures of spread," which "does
 211 not require the assumption of symmetry." From that, a popular
 212 efficient scale estimator, the Rousseeuw-Croux scale estima-
 213 tor (20), was derived in 1993, but the importance of tackling
 214 the symmetry assumption has been greatly underestimated.
 215 While they had already considered one version of the trimmed
 216 standard deviation in the third paper of that series (21), in
 217 the final section of that paper (19), they explored another two
 218 possible versions, which were modified here for comparison,

$$\left[n\left(\frac{1}{2} - \epsilon\right)\right]^{-\frac{1}{2}} \left[\sum_{i=\frac{n}{2}}^{n(1-\epsilon)} [X_i - X_{n-i+1}]^2 \right]^{\frac{1}{2}}, \quad [1] \quad 219$$

220 and

$$\left[\binom{n}{2} (1 - \epsilon - \gamma\epsilon)\right]^{-\frac{1}{2}} \left[\sum_{i=\binom{n}{2}\gamma\epsilon}^{\binom{n}{2}(1-\epsilon)} (X - X')_i^2 \right]^{\frac{1}{2}}, \quad [2] \quad 221$$

222 where $(X - X')_1 \leq \dots \leq (X - X')_{\binom{n}{2}}$ are the order statistics
 223 of the "pseudo-sample", $X_i - X_j$, $i < j$. The paper ended
 224 with, "We do not know a fortiori which of the measures [1] or
 225 [2] is preferable and leave these interesting questions open."

226 Observe that the kernel of the unbiased estimation of the
 227 second central moment by using U -statistic is $\psi_2(x_1, x_2) =$
 228 $\frac{1}{2}(x_1 - x_2)^2$. If adding the $\frac{1}{2}$ term in [2], as $\epsilon \rightarrow 0$, the result
 229 is equivalent to the standard deviation estimated by using
 230 U -statistic (also noted by Janssen, Serfling, and Veraverbeke
 231 in 1987) (22). In fact, they also showed that, when ϵ is 0, [2]
 232 is $\sqrt{2}$ times the standard deviation.

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

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

249 In the previous article on semiparametric robust mean es-
 250 timation, it was shown that the bias of a robust estimator
 251 with an adjustable breakdown point is often monotonic with
 252 respect to the breakdown point in a semiparametric distribu-
 253 tion. Naturally, the estimator's name should correspond to
 254 the population parameter with which it is consistent as $\epsilon \rightarrow 0$.
 255 The trimmed standard deviation following this nomenclature

256 is $Tsd_{\epsilon_U=1-\sqrt{1-\epsilon}, \gamma, n} := \left[TM_{\epsilon, \gamma} \left((\psi_2(X_{N_1}, X_{N_2}))_{N=1}^{\binom{n}{2}} \right) \right]^{-\frac{1}{2}}$,
 257 where $TM_{\epsilon, \gamma}(Y)$ denotes the ϵ, γ -trimmed mean with the se-
 258 quence $(\psi_2(X_{N_1}, X_{N_2}))_{N=1}^{\binom{n}{2}}$ as an input, the proof of the break-
 259 down point is given in Subsection ???. Removing the square

root yields the trimmed variance ($\text{Tvar}_{\epsilon U_2, \gamma, n}$). 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. \square

Previously, it was shown that any symmetric distribution with a finite second moment is ν th 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, unbiasedness 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 k th 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-Lehmann 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 k th central moment ($k \leq n$) is defined as,

$$\text{Wkm}_{\epsilon U_k, \gamma, n} := \text{WL}_{\epsilon, \gamma, n} \left((\psi_k(X_{N_1}, \dots, X_{N_k}))_{N=1}^{\binom{n}{k}} \right),$$

where $\epsilon U_k = 1 - (1 - \epsilon)^{\frac{1}{k}}$, X_{N_1}, \dots, X_{N_k} are the n choose k elements from X , $\psi_k(x_1, \dots, x_k) = \sum_{j=0}^{k-2} (-1)^j \left(\frac{1}{k-j} \right) \sum (x_{i_1}^{k-j} \dots x_{i_{j+1}}) + (-1)^{k-1} (k-1) x_1 \dots x_k$, the second summation is over

$i_1, \dots, i_{j+1} = 1$ to k with $i_1 < \dots < i_{j+1}$ (30). Despite the complexity, the structure of the k th central moment kernel distributions can be elucidated by decomposing.

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

Proof. Without loss of generality, generating the distribution of the k -tuple $(Q(p_{i_1}), \dots, Q(p_{i_k}))$ under continuity, $k > 2$, $i_1 < \dots < i_k$, $p_{i_1} < \dots < p_{i_k}$, the corresponding probability density is $f_{X, \dots, X}(Q(p_{i_1}), \dots, Q(p_{i_k})) = k! f(Q(p_{i_1})) \dots f(Q(p_{i_k}))$. Transforming the distribution of the k -tuple by the function $\psi_k(x_1, \dots, x_k)$, denoting $\bar{\Delta} = \psi_k(Q(p_{i_1}), \dots, Q(p_{i_k}))$. The probability $f_{\Xi_k}(\bar{\Delta}) = \sum_{\bar{\Delta} = \psi_k(Q(p_{i_1}), \dots, Q(p_{i_k}))} f_{X, \dots, X}(Q(p_{i_1}), \dots, Q(p_{i_k}))$ is the summation of the probabilities of all k -tuples such that $\bar{\Delta}$ is equal to $\psi_k(Q(p_{i_1}), \dots, Q(p_{i_k}))$. 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, \dots, x_{k-1} exhaust all combinations under the inequality constraints, i.e., $Q(p_{i_1}) = x_1 < x_2 < \dots < x_{k-1} < x_k = Q(p_{i_k})$. 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 $x_1 = \dots = x_k = 0$, where $\psi_k = 0$. Other candidates are within the boundaries, i.e., $\psi_k(x_1 = x_1, x_2 = x_k, \dots, x_k = x_k)$, \dots , $\psi_k(x_1 = x_1, \dots, x_i = x_1, x_{i+1} = x_k, \dots, x_k = x_k)$, \dots , $\psi_k(x_1 = x_1, \dots, x_{k-1} = x_1, x_k = x_k)$. $\psi_k(x_1 = x_1, \dots, x_i = x_1, x_{i+1} = x_k, \dots, x_k = x_k)$ 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 g th group has $i \binom{i-1}{g-j-1} \binom{k-i}{j}$ terms having the form $(-1)^{g+1} \frac{1}{k-g+1} x_1^{k-j} x_k^j$, from $k-j+1$ th to $i+j$ th group, the g th group has $i \binom{i-1}{g-j-1} \binom{k-i}{j} + (k-i) \binom{k-i-1}{j-k+g-1} \binom{i}{k-j}$ terms having the form $(-1)^{g+1} \frac{1}{k-g+1} x_1^{k-j} x_k^j$. If $j < \frac{k+1-i}{2}$, from $j+1$ st to $i+j$ th group, the g th group has $i \binom{i-1}{g-j-1} \binom{k-i}{j}$ terms having the form $(-1)^{g+1} \frac{1}{k-g+1} x_1^{k-j} x_k^j$. If $j \geq \frac{k}{2}$, from $k-j+1$ st to j th group, the g th group has $(k-i) \binom{k-i-1}{j-k+g-1} \binom{i}{k-j}$ terms having the form $(-1)^{g+1} \frac{1}{k-g+1} x_1^{k-j} x_k^j$, from $j+1$ th to $j+i$ th group, $i+j < k$, the g th group has $i \binom{i-1}{g-j-1} \binom{k-i}{j} + (k-i) \binom{k-i-1}{j-k+g-1} \binom{i}{k-j}$ terms having the form $(-1)^{g+1} \frac{1}{k-g+1} x_1^{k-j} x_k^j$. The final k th group is the term $(-1)^{k-1} (k-1) x_1^{k-i} x_k^i$. 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 $(-1)^{k-1} (k-1) + \sum_{g=i+1}^{k-1} (-1)^{g+1} \frac{1}{k-g+1} (k-i) \binom{k-i-1}{g-i-1} + \sum_{g=k-i+1}^{k-1} (-1)^{g+1} \frac{1}{k-g+1} i \binom{i-1}{g-k+i-1} = (-1)^{k-1} (k-1) + (-1)^{k+1} + (k-i) (-1)^k + (-1)^k (i-1) = (-1)^{k+1}$. The summation identities are $\sum_{g=i+1}^{k-1} (-1)^{g+1} \frac{1}{k-g+1} (k-i) \binom{k-i-1}{g-i-1} = (k-i) \int_0^1 \sum_{g=i+1}^{k-1} (-1)^{g+1} \binom{k-i-1}{g-i-1} t^{k-g} dt$ $(k-i) \int_0^1 ((-1)^i (t-1)^{k-i-1} - (-1)^{k+1}) dt$

$(k-i) \binom{(-1)^k}{i-k} + (-1)^k = (-1)^{k+1} + (k-i)(-1)^k$
 and $\sum_{g=k-i+1}^{k-1} (-1)^{g+1} \frac{1}{k-g+1} i \binom{i-1}{g-k+i-1} =$
 $\int_0^1 \sum_{g=k-i+1}^{k-1} (-1)^{g+1} i \binom{i-1}{g-k+i-1} t^{k-g} dt =$
 $\int_0^1 (i(-1)^{k-i} (t-1)^{i-1} - i(-1)^{k+1}) dt = (-1)^k (i-1).$
 If $j < \frac{k+1-i}{2}$, $i > k-1$, if $i = k$, $\psi_k = 0$, if $\frac{k+1-i}{2} \leq j \leq \frac{k-1}{2}$,
 $\frac{k+1}{2} \leq i \leq k-1$, 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 co-
 efficient 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. There-
 fore, $\psi_k(x_1 = x_1, \dots, x_i = x_1, x_{i+1} = x_k, \dots, x_k = x_k) =$
 $\binom{k}{i}^{-1} (-1)^{1+i} (x_1 - x_k)^k$, the maximum and minimum of ψ_k
 follow directly from the properties of the binomial coeffi-
 cient. \square

ξ_Δ is closely related to $f_\Xi(\Delta)$, which is the pairwise differ-
 ence distribution, since the probability density of ξ_Δ can be ex-
 pressed as $f_{\Xi_k}(\bar{\Delta}|\Delta) = \sum_{\bar{\Delta} = -(\frac{k}{3+(-1)^k})}^{\frac{1}{k}(-\Delta)^k} (-1)^k f_{\Xi_k}(\bar{\Delta}|\Delta) =$
 $f_\Xi(\Delta) = \int_0^\infty 2f(t)f(t-\Delta)dt$. 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 medi-
 ans 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 $\leq 0.1\sigma$ 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 aver-
 age, 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 $\frac{\partial \psi_3(x_1, x_2, x_3)}{\partial x_2} = -\frac{x_1^2}{2} - x_1 x_2 + 2x_1 x_3 + x_2^2 - x_2 x_3 - \frac{x_3^2}{2}$,
 $-\frac{3}{4}(x_1 - x_3)^2 \leq \frac{\partial \psi_3(x_1, x_2, x_3)}{\partial x_2} \leq -\frac{1}{2}(x_1 - x_3)^2 \leq 0$. 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 (the
 same result can be inferred from the definition of central mo-
 ments, 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 k th
 central moment kernel distribution, $k > 2$, forcing it to be
 unimodal-like with the mode and median close to zero, then,
 the inequality $f(Q(\epsilon)) \leq f(Q(1-\epsilon))$ or $f(Q(\epsilon)) \geq f(Q(1-\epsilon))$
 holds in general. If a distribution is ordered and all of its cen-
 tral 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 dis-
 tribution, location invariant, is introduced in the next theorem.
 The proof is provided in the SI Text.

Theorem B.3. $\psi_k(x_1 = \lambda x_1 + \mu, \dots, x_k = \lambda x_k + \mu) =$
 $\lambda^k \psi_k(x_1, \dots, x_k).$

A direct result of Theorem B.3 is that, Wkm after stan-
 dardization is invariant to location and scale. So, the weighted
 standardized k th moment is defined to be

$$Wskm_{\epsilon_{U_k}, \gamma, n} := \frac{Wkm_{\epsilon_{U_k}, \gamma, n}}{Wvar_{\epsilon_{U_2}, \gamma, n}^{\frac{k}{2}}}.$$

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

$$rkm_{d, \epsilon_{U_k}, n} := (d+1)SWkm_{\epsilon_{U_k}, n} - dmkm_n,$$

where mkm_n is the median k th central moment. Since
 $SWkm_{\epsilon_{U_k}, n}$ 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 k th central moment is thus defined as

$$qkm_{d, \epsilon_{U_k}, n} := \hat{Q}_n \left(\left(pSWkm_{\epsilon_{U_k}, n} - \frac{1}{2} \right) d + pSWkm_{\epsilon_{U_k}, n} \right),$$

where $pSWkm_{\epsilon_{U_k}, n} = \hat{F}_{\psi, n}(SWkm_{\epsilon_{U_k}, n})$, $\hat{F}_{\psi, n}$ is the em-
 pirical 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 quan-
 tile kurtosis are defined to be $qskew_{d, \epsilon_{U_3}, n} := \frac{qtm_{d, \epsilon_{U_3}, n}}{qsd_{d, \epsilon_{U_2}, n}^3}$

and $qkurt_{d, \epsilon_{U_4}, n} := \frac{qfm_{d, \epsilon_{U_4}, n}}{qsd_{d, \epsilon_{U_2}, n}^4}$. Quantile standard deviation

$(qsd_{d, \epsilon_{U_2}, n})$, recombined standard deviation ($rsd_{d, \epsilon_{U_2}, n}$), quan-
 tile third central moment ($qtm_{d, \epsilon_{U_3}, n}$), quantile fourth cen-
 tral 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, $m = \lambda \sqrt[\alpha]{\ln(2)}$, $\mu = \lambda \Gamma(1 + \frac{1}{\alpha})$, 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 $QA(\epsilon, \gamma, \alpha_1, \dots, \alpha_i, \dots, \alpha_k)$, where α_i represent the parameters of the distribution, then, a distribution is γ -congruent if and only if the sign of $\frac{\partial QA}{\partial \alpha_i}$ remains the same for all $0 \leq \epsilon \leq \frac{1}{1+\gamma}$. 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 $\frac{\partial QA}{\partial \alpha_i}$ remains the same for all $0 \leq \epsilon \leq \frac{1}{1+\gamma}$ 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. \square

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 $\lambda WA_0(\epsilon) + \mu$, where $WA_0(\epsilon)$ is an integral of $Q_0(p)$ 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. \square

Theorem C.3. *The second central moment kernel 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. \square

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 $\frac{\partial QA(\epsilon, \gamma, \alpha)}{\partial \alpha} < 0$, the Pareto distribution is γ -congruent. The derivative for the lognormal distribution is $\frac{\partial SQA(\epsilon, \sigma)}{\partial \sigma} = \frac{-\text{erfc}^{-1}(2\epsilon)e^{\mu - \sqrt{2}\sigma \text{erfc}^{-1}(2\epsilon)} - \text{erfc}^{-1}(2-2\epsilon)e^{\mu - \sqrt{2}\sigma \text{erfc}^{-1}(2-2\epsilon)}}{\sqrt{2}}$. Since the inverse complementary error function is positive when the input is smaller than 1, and negative when the input is larger than 1, $\text{erfc}^{-1}(2\epsilon) = -\text{erfc}^{-1}(2-2\epsilon)$, $e^{\mu - \sqrt{2}\sigma \text{erfc}^{-1}(2-2\epsilon)} > e^{\mu - \sqrt{2}\sigma \text{erfc}^{-1}(2\epsilon)}$, $\frac{\partial SQA(\epsilon, \sigma)}{\partial \sigma} > 0$, 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 invariant moments and their related weighted moments are given, there should exist a unique k -tuple (d_{im}, \dots, 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

$$\begin{cases} rm(SWL, m, D(|rskew|, rkurt, 1)) = \mu \\ rvar(SWvar, mvar, D(|rskew|, rkurt, 2)) = \mu_2 \\ rtm(SWtm, mtm, D(|rskew|, rkurt, 3)) = \mu_3 \\ rf m(SWfm, mfm, D(|rskew|, rkurt, 4)) = \mu_4 \\ rskew = \frac{\mu_3}{\mu_2} \\ rkurt = \frac{\mu_4}{\mu_2^2} \end{cases},$$

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 $\varsigma(|rskew|) = \left| \frac{rtm(SWtm, mtm, D(|rskew|, 3))}{rvar(SWvar, mvar, D(|rskew|, 2))^{\frac{3}{2}}} \right|$ and $\varkappa(rkurt) = \frac{rfm(SWfm, mfm, D(rkurt, 4))}{rvar(SWvar, mvar, D(rkurt, 2))^2}$. 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: $rkurt_{i-1}$

```

1:  $i = 0$ 
2:  $rkurt_i \leftarrow \varkappa(kurtosis_{max})$   $\triangleright$  Using the maximum kurtosis available in  $D$  as an initial guess.
3: repeat
4:    $i = i + 1$ 
5:    $rkurt_{i-1} \leftarrow rkurt_i$ 
6:    $rkurt_i \leftarrow \varkappa(rkurt_{i-1})$ 
7: until  $i > maxit$  or  $|rkurt_i - rkurt_{i-1}| < \delta$   $\triangleright maxit$  is the maximum number of iterations,  $\delta$  is a small positive number.
```

The following theorem shows the validity 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, 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 assumed to be 1. From the definition of D , $\frac{\varkappa(rkurt_D)}{rkurt_D} = \frac{\frac{fm_D - SWfm_D}{SWfm_D - mfm_D} (SWfm - mfm) + SWfm}{rkurt_D \left(\frac{var_D - SWvar_D}{SWvar_D - mvar_D} (SWvar - mvar) + SWvar \right)^2}$, 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 $mkms$ are all equal to zero, $\frac{\varkappa(rkurt_D)}{rkurt_D} = \frac{\frac{fm_D - SWfm_D}{SWfm_D} (SWfm) + SWfm}{rkurt_D \left(\frac{SWvar_D}{SWvar_D} \right)^2} = \frac{\left(\frac{fm_D - SWfm_D}{SWfm_D} + 1 \right) (SWfm)}{fm_D \left(\frac{SWvar_D}{SWvar_D} \right)^2} = \frac{SWfm SWvar_D^2}{SWfm_D SWvar_D^2} = \frac{SWkurt}{SWkurt_D}$. Since $SWfm_D$ 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, $SWkurt$ is a measure of kurtosis that is invariant to location and scale, so $\lim_{rkurt_D \rightarrow \infty} \frac{\varkappa(rkurt_D)}{rkurt_D} < 1$. As a result, if there is at least one fix point, let the largest one be

666 fix_{max} , then it is attracting since $|\frac{\partial(\kappa(rkurt_D))}{\partial(rkurt_D)}| < 1$ for all
 667 $rkurt_D \in [fix_{max}, kurtosis_{max}]$.

668 Consider any $SWkurt_D > SWkurt$, $\frac{\kappa(rkurt_D)}{rkurt_D} < 1$, the
 669 same logic applies, a consistent estimator must be the last
 670 attracting fix point, so fix_{max} is consistent. \square

671 As a result of Theorem D.1, assuming continuity, $mkms$
 672 are all equal to zero, and congruence of the central moment
 673 kernel distributions, Algorithm 1 converges surely provided
 674 that a fix point exists within the domain of D . At this stage,
 675 D can only be approximated through a Monte Carlo study.
 676 Continuity can be ensured by using linear interpolation. One
 677 common encountered problem is that the domain of D de-
 678 pends on both the consistent distribution and the Monte Carlo
 679 study, so the iteration may halt at the boundary if the fix
 680 point is not within the domain. However, by setting a proper
 681 maximum number of iterations, the algorithm can return the
 682 optimal boundary value. For quantile moments, the logic is
 683 similar, if the percentiles do not exceed the breakdown point.
 684 If this is the case, consistent estimation is impossible, and
 685 the algorithm will stop due to the maximum number of iter-
 686 ations. The fix point iteration is, in principle, similar to the
 687 iterative reweighing in M-estimator, but an advantage of this
 688 algorithm is that the optimization is solely related to the d
 689 value function and is independent of the sample size (except
 690 for the quantile moments, which require re-computation of the
 691 quantile function, but this operation has a time complexity of
 692 $O(1)$ for a sorted sample). Since $|rskew|$ can specify d_{rm} after
 693 optimization, this algorithm enables the robust estimations of
 694 all four moments to reach a near-consistent level for common
 695 unimodal distributions (Table ??, SI Dataset S1), just using
 696 the Weibull distribution as the consistent distribution.

697 **Data Availability.** Data for Table ?? are given in SI Dataset S1.
 698 All codes have been deposited in [GitHub](#).

699 **ACKNOWLEDGMENTS.** I gratefully acknowledge the construc-
 700 tive comments made by the editor which substantially improved
 701 the clarity and quality of this paper.

- 702 1. CF Gauss, *Theoria combinationis observationum erroribus minimis obnoxiae*. (Henricus
 703 Dieterich), (1823).
- 704 2. S Newcomb, A generalized theory of the combination of observations so as to obtain the best
 705 result. *Am. journal Math.* **8**, 343–366 (1886).
- 706 3. S Newcomb, Researches on the motion of the moon. part ii, the mean motion of the moon and
 707 other astronomical elements derived from observations of eclipses and occultations extending
 708 from the period of the babylonians until ad 1908. *United States. Naut. Alm. Off. Astron. paper*;
 709 v. 9 **9**, 1 (1912).
- 710 4. PJ Huber, Robust estimation of a location parameter. *Ann. Math. Stat.* **35**, 73–101 (1964).
- 711 5. X He, WK Fung, Method of medians for lifetime data with weibull models. *Stat. medicine* **18**,
 712 1993–2009 (1999).
- 713 6. M Menon, Estimation of the shape and scale parameters of the weibull distribution. *Techno-*
 714 *metrics* **5**, 175–182 (1963).
- 715 7. SD Dubey, Some percentile estimators for weibull parameters. *Technometrics* **9**, 119–129
 716 (1967).
- 717 8. KM Hassanein, Percentile estimators for the parameters of the weibull distribution. *Biometrika*
 718 **58**, 673–676 (1971).
- 719 9. NB Marks, Estimation of weibull parameters from common percentiles. *J. applied Stat.* **32**,
 720 17–24 (2005).
- 721 10. K Boudt, D Caliskan, C Croux, Robust explicit estimators of weibull parameters. *Metrika* **73**,
 722 187–209 (2011).
- 723 11. SD Dubey, *Contributions to statistical theory of life testing and reliability*. (Michigan State
 724 University of Agriculture and Applied Science. Department of statistics), (1960).
- 725 12. LJ Bain, CE Antle, Estimation of parameters in the weibull distribution. *Technometrics* **9**,
 726 621–627 (1967).
- 727 13. RV Hogg, Adaptive robust procedures: A partial review and some suggestions for future
 728 applications and theory. *J. Am. Stat. Assoc.* **69**, 909–923 (1974).
- 729 14. RJ Hyndman, Y Fan, Sample quantiles in statistical packages. *The Am. Stat.* **50**, 361–365
 730 (1996).
- 731 15. C Bernard, R Kazzi, S Vanduffel, Range value-at-risk bounds for unimodal distributions under
 732 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. associa-*
tion **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 Cléménçon, P Bertail, On medians of (randomized) pairwise means in *Interna-*
tional 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).