Robust estimations from distribution structures: II. Central Moments

Tuobang Li

This manuscript was compiled on November 29, 2023

In 1954, Hodges and Lehmann demonstrated that if X and Y are independently sampled from an identical unimodal distribution, $X-Y$ will exhibit symmetrical unimodality with its peak centered at zero. Building upon this foundational work, the current study delves into the structure of the kernel distribution of U-statistics. It is shown that the $\mathbf{k}$th central moment kernel distributions ($\mathbf{k}>2$) derived from a unimodal distribution exhibit location invariance and is also nearly unimodal with the mode and median close to zero. This article provides an approach to study the general structure of kernel distributions.

moments | invariant | unimodal | U-statistics

The most popular robust scale estimator currently, the median absolute deviation, was popularized by Hampel (1974) (1), who credits the idea to Gauss in 1816 (2). In 1976, in their landmark series *Descriptive Statistics for Nonparametric Models*, Bickel and Lehmann (3) generalized a class of estimators as measures of the dispersion of a symmetric distribution around its center of symmetry. In 1979, the same series, they (4) proposed a class of estimators referred to as measures of spread, which consider the pairwise differences of a random variable, irrespective of its symmetry, throughout its distribution, rather than focusing on dispersion relative to a fixed point. In the final section (4), they explored a version of the trimmed standard deviation based on pairwise differences, which is modified here for comparison,

$$\left[\binom{n}{2}\left(1-\epsilon_{\mathbf{0}}-\gamma\epsilon_{\mathbf{0}}\right)\right]^{-\frac{1}{2}}\left[\sum_{i=\binom{n}{2}\gamma\epsilon_{\mathbf{0}}}^{\binom{n}{2}(1-\epsilon_{\mathbf{0}})}\left(X_{i_1}-X_{i_2}\right)_i^{2}\right]^{\frac{1}{2}}, \qquad [1]$$

where $(X_{i_1}-X_{i_2})_1 \leq \ldots \leq (X_{i_1}-X_{i_2})_{\binom{n}{2}}$ are the order statistics of $X_{i_1}-X_{i_2}$, $i_1<i_2$, provided that $\binom{n}{2}\gamma\epsilon_{\mathbf{0}} \in \mathbb{N}$ and $\binom{n}{2}(1-\epsilon_{\mathbf{0}}) \in \mathbb{N}$. They showed that, when $\epsilon_{\mathbf{0}}=0$, the result obtained using [1] is equal to $\sqrt{2}$ times the sample standard deviation. The paper ended with, "We do not know a fortiori which of the measures is preferable and leave these interesting questions open."

Two examples of the impacts of that series are as follows. Oja (1981, 1983) (5, 6) provided a more comprehensive and generalized examination of these concepts, and integrated the measures of location, dispersion, and spread as proposed by Bickel and Lehmann (3, 4, 7), along with van Zwet's convex transformation order of skewness and kurtosis (1964) (8) for univariate and multivariate distributions, resulting a greater degree of generality and a broader perspective on these statistical constructs. Rousseeuw and Croux proposed a popular efficient scale estimator based on separate medians of pairwise differences taken over i_1 and i_2 (9) in 1993. However the importance of tackling the symmetry assumption has been greatly underestimated, as will be discussed later.

To address their open question (4), the nomenclature used in this paper is introduced as follows:

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

In REDS I, it was shown that the bias of a robust estimator with an adjustable breakdown point is often monotonic with respect to the breakdown point in a semiparametric distribution. Naturally, the estimator's name should reflect the population parameter that it approaches as $\epsilon\to 0$. If multiplying all pseudo-samples by a factor of $\frac{1}{\sqrt{2}}$, then [1] is the trimmed standard deviation adhering to this nomenclature, since $\psi_2(x_1,x_2)=\frac{1}{2}(x_1-x_2)^2$ is the kernel function of the unbiased estimation of the second central moment by using U-statistic (10). This definition should be preferable, not only because it is the square root of a trimmed U-statistic, which is closely related to the minimum-variance unbiased estimator (MVUE), but also because the second γ-orderliness of the second central moment kernel distribution is ensured by the next exciting theorem.

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

Proof. In 1954, Hodges and Lehmann established that if X and Y are independently drawn from the same unimodal distribution, $X-Y$ will be a symmetric unimodal distribution peaking at zero (11). Given the constraint in the pairwise differences

Significance Statement

In nonparametric statistics, the focus is on the relative differences of robust estimators, which is considered more crucial than their precise values. This principle implies that if the underlying distribution's parameters shift, then all corresponding nonparametric estimates, provided they target the same characteristic of the distribution, are expected to uniformly and asymptotically adjust in a consistent direction. This article discusses the validity of this fundamental principle of nonparametrics in various scenarios. It is found that for the $\mathbf{k}$th central moment, kernel distributions generally follow this principle.

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

that $X_{i_1} < X_{i_2}$, $i_1 < i_2$, it directly follows from Theorem 1 in (11) that the pairwise difference distribution (Ξ_Δ) generated from any unimodal distribution is always monotonic increasing with a mode at zero. Since $X - X'$ is a negative variable that is monotonically increasing, applying the squaring transformation, the relationship between the original variable $X - X'$ and its squared counterpart $(X - X')^2$ can be represented as follows: $X - X' < Y - Y' \implies (X - X')^2 > (Y - Y')^2$. In other words, as the negative values of $X - X'$ become larger in magnitude (more negative), their squared values $(X - X')^2$ become larger as well, but in a monotonically decreasing manner with a mode at zero. Further multiplication by $\frac{1}{2}$ also does not change the monotonicity and mode, since the mode is zero. Therefore, the transformed pdf becomes monotonically decreasing with a mode at zero. In REDS I, it was proven that a right-skewed distribution with a monotonic decreasing pdf is always second γ-ordered, which gives the desired result. ☐

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

On the other hand, while robust estimation of scale has been intensively studied with established methods (3, 4), the development of robust measures of asymmetry and kurtosis lags behind, despite the availability of several approaches (12–16). 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 define a convenient approach to quantitatively estimate the estimators' efficiencies.

Robust Estimations of the Central Moments

In 1928, Fisher constructed $\mathbf{k}$-statistics as unbiased estimators of cumulants (17). Halmos (1946) proved that a functional θ admits an unbiased estimator if and only if it is a regular statistical functional of degree $\mathbf{k}$ and showed a relation of symmetry, unbiasness and minimum variance (18). Hoeffding, in 1948, generalized U-statistics (19) which enable the derivation of a minimum-variance unbiased estimator from each unbiased estimator of an estimable parameter. In 1984, Serfling pointed out the speciality of Hodges-Lehmann estimator, which is neither a simple L-statistic nor a U-statistic, and considered the generalized L-statistics and trimmed U-statistics (20). Given a kernel function $h_\mathbf{k}$ which is a symmetric function of $\mathbf{k}$ variables, the LU-statistic is defined as:

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

where $\epsilon = 1 - (1-\epsilon_\mathbf{0})^{\frac{1}{\mathbf{k}}}$ (proven in Subsection ??), $X_{N_1},\ldots,X_{N_\mathbf{k}}$ are the n choose $\mathbf{k}$ elements from the sample, $LL_{k,\epsilon_\mathbf{0},\gamma,n}(Y)$ denotes the LL-statistic with the sorted sequence $\text{sort}\left(\left(h_\mathbf{k}\left(X_{N_1},\ldots,X_{N_\mathbf{k}}\right)\right)_{N=1}^{\binom{n}{\mathbf{k}}}\right)$ serving as an input. In the context of Serfling's work, the term 'trimmed U-statistic' is used when $LL_{k,\epsilon_\mathbf{0},\gamma,n}$ is $\text{TM}_{\epsilon_\mathbf{0},\gamma,n}$ (20).

In 1997, Heffernan (10) obtained an unbiased estimator of the $\mathbf{k}$th central moment by using U-statistics and demonstrated that it is the minimum variance unbiased estimator for distributions with the finite first $\mathbf{k}$ moments. The weighted H-L $\mathbf{k}$th central moment ($2 \le \mathbf{k} \le n$) is thus defined as,

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

where $\text{WHLM}_{k,\epsilon_\mathbf{0},\gamma,n}$ is used as the $LL_{k,\epsilon_\mathbf{0},\gamma,n}$ in LU, $\psi_\mathbf{k}(x_1,\ldots,x_\mathbf{k}) = \sum_{j=0}^{\mathbf{k}-2}(-1)^j\left(\frac{1}{\mathbf{k}-j}\right)\sum\left(x_{i_1}^{\mathbf{k}-j}x_{i_2}\ldots x_{i_{j+1}}\right) + (-1)^{\mathbf{k}-1}(\mathbf{k}-1)\,x_1\ldots x_\mathbf{k}$, the second summation is over $i_1,\ldots,i_{j+1} = 1$ to $\mathbf{k}$ with $i_1 \ne i_2 \ne \ldots \ne i_{j+1}$ and $i_2 < i_3 < \ldots < i_{j+1}$ (10). Despite the complexity, the following theorem offers an approach to infer the general structure of such kernel distributions.

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

Proof. The support of ξ_Δ is the extrema of the function $\psi_\mathbf{k}(Q(p_1),\cdots,Q(p_\mathbf{k}))$ subjected to the constraints, $Q(p_1) < \cdots < Q(p_\mathbf{k})$ and $\Delta = Q(p_1) - Q(p_\mathbf{k})$. Using the Lagrange multiplier, the only critical point can be determined at $Q(p_1) = \cdots = Q(p_\mathbf{k}) = 0$, where $\psi_\mathbf{k} = 0$. Other candidates are within the boundaries, i.e., $\psi_\mathbf{k}(x_1 = Q(p_1), x_2 = Q(p_\mathbf{k}), \cdots, x_\mathbf{k} = Q(p_\mathbf{k}))$, $\cdots$, $\psi_\mathbf{k}(x_1 = Q(p_1), \cdots, x_i = Q(p_1), x_{i+1} = Q(p_\mathbf{k}), \cdots, x_\mathbf{k} = Q(p_\mathbf{k}))$, $\cdots$, $\psi_\mathbf{k}(x_1 = Q(p_1), \cdots, x_{\mathbf{k}-1} = Q(p_1), x_\mathbf{k} = Q(p_\mathbf{k}))$. $\psi_\mathbf{k}(x_1 = Q(p_1), \cdots, x_i = Q(p_1), x_{i+1} = Q(p_\mathbf{k}), \cdots, x_\mathbf{k} = Q(p_\mathbf{k}))$ can be divided into $\mathbf{k}$ groups. The gth group has the common factor $(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}$, if $1 \le g \le \mathbf{k}-1$ and the final $\mathbf{k}$th group is the term $(-1)^{\mathbf{k}-1}(\mathbf{k}-1)\,Q(p_1)^iQ(p_\mathbf{k})^{\mathbf{k}-i}$. If $\frac{\mathbf{k}+1-i}{2} \le j \le \frac{\mathbf{k}-1}{2}$ and $j+1 \le g \le \mathbf{k}-j$, the gth group has $i\binom{i-1}{g-j-1}\binom{\mathbf{k}-i}{j}$ terms having the form $(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}Q(p_1)^{\mathbf{k}-j}Q(p_\mathbf{k})^j$. If $\frac{\mathbf{k}+1-i}{2} \le j \le \frac{\mathbf{k}-1}{2}$ and $\mathbf{k}-j+1 \le g \le i+j$, the gth group has $i\binom{i-1}{g-j-1}\binom{\mathbf{k}-i}{j} + (\mathbf{k}-i)\binom{\mathbf{k}-i-1}{j-\mathbf{k}+g-1}\binom{i}{\mathbf{k}-j}$ terms having the form $(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}Q(p_1)^{\mathbf{k}-j}Q(p_\mathbf{k})^j$. If $0 \le j < \frac{\mathbf{k}+1-i}{2}$ and $j+1 \le g \le i+j$, the gth group has $i\binom{i-1}{g-j-1}\binom{\mathbf{k}-i}{j}$ terms having the form $(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}Q(p_1)^{\mathbf{k}-j}Q(p_\mathbf{k})^j$. If $\frac{\mathbf{k}}{2} \le j \le \mathbf{k}$ and $\mathbf{k}-j+1 \le g \le j$, the gth group has $(\mathbf{k}-i)\binom{\mathbf{k}-i-1}{j-\mathbf{k}+g-1}\binom{i}{\mathbf{k}-j}$ terms having the form $(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}Q(p_1)^{\mathbf{k}-j}Q(p_\mathbf{k})^j$. If $\frac{\mathbf{k}}{2} \le j \le \mathbf{k}$ and $j+1 \le g \le j+i < \mathbf{k}$, the gth group has

$i\binom{i-1}{g-j-1}\binom{\mathbf{k}-i}{j}+(\mathbf{k}-i)\binom{\mathbf{k}-i-1}{j-\mathbf{k}+g-1}\binom{i}{\mathbf{k}-j}$ terms having the form $(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}Q(p_1)^{\mathbf{k}-j}Q(p_\mathbf{k})^j$. So, if $i+j=\mathbf{k}$, $\frac{\mathbf{k}}{2}\le j\le \mathbf{k}$, $0\le i\le \frac{\mathbf{k}}{2}$, the summed coefficient of $Q(p_1)^iQ(p_\mathbf{k})^{\mathbf{k}-i}$ is $(-1)^{\mathbf{k}-1}(\mathbf{k}-1)+\sum_{g=i+1}^{\mathbf{k}-1}(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}(\mathbf{k}-i)\binom{\mathbf{k}-i-1}{g-i-1}+\sum_{g=\mathbf{k}-i+1}^{\mathbf{k}-1}(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}i\binom{i-1}{g-\mathbf{k}+i-1}=(-1)^{\mathbf{k}-1}(\mathbf{k}-1)+(-1)^{\mathbf{k}+1}+(\mathbf{k}-i)(-1)^{\mathbf{k}}+(-1)^{\mathbf{k}}(i-1)=(-1)^{\mathbf{k}+1}$. The summation identities are $\sum_{g=i+1}^{\mathbf{k}-1}(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}(\mathbf{k}-i)\binom{\mathbf{k}-i-1}{g-i-1}=(\mathbf{k}-i)\int_0^1\sum_{g=i+1}^{\mathbf{k}-1}(-1)^{g+1}\binom{\mathbf{k}-i-1}{g-i-1}t^{\mathbf{k}-g}dt=(\mathbf{k}-i)\int_0^1\left((-1)^i(t-1)^{\mathbf{k}-i-1}-(-1)^{\mathbf{k}+1}\right)dt=(\mathbf{k}-i)\left(\frac{(-1)^{\mathbf{k}}}{i-\mathbf{k}}+(-1)^{\mathbf{k}}\right)=(-1)^{\mathbf{k}+1}+(\mathbf{k}-i)(-1)^{\mathbf{k}}$ and $\sum_{g=\mathbf{k}-i+1}^{\mathbf{k}-1}(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}i\binom{i-1}{g-\mathbf{k}+i-1}=\int_0^1\sum_{g=\mathbf{k}-i+1}^{\mathbf{k}-1}(-1)^{g+1}i\binom{i-1}{g-\mathbf{k}+i-1}t^{\mathbf{k}-g}dt=\int_0^1\left(i(-1)^{\mathbf{k}-i}(t-1)^{i-1}-i(-1)^{\mathbf{k}+1}\right)dt=(-1)^{\mathbf{k}}(i-1)$. If $0\le j<\frac{\mathbf{k}+1-i}{2}$ and $i=\mathbf{k}$, $\psi_\mathbf{k}=0$. If $\frac{\mathbf{k}+1-i}{2}\le j\le\frac{\mathbf{k}-1}{2}$ and $\frac{\mathbf{k}+1}{2}\le i\le \mathbf{k}-1$, the summed coefficient of $Q(p_1)^iQ(p_\mathbf{k})^{\mathbf{k}-i}$ is $(-1)^{\mathbf{k}-1}(\mathbf{k}-1)+\sum_{g=\mathbf{k}-i+1}^{\mathbf{k}-1}(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}i\binom{i-1}{g-\mathbf{k}+i-1}+\sum_{g=i+1}^{\mathbf{k}-1}(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}(\mathbf{k}-i)\binom{\mathbf{k}-i-1}{g-i-1}$, the same as above. If $i+j<\mathbf{k}$, since $\binom{i}{\mathbf{k}-j}=0$, the related terms can be ignored, so, using the binomial theorem and beta function, the summed coefficient of $Q(p_1)^{k-j}Q(p_\mathbf{k})^j$ is $\sum_{g=j+1}^{i+j}(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}i\binom{i-1}{g-j-1}\binom{\mathbf{k}-i}{j}=i\binom{\mathbf{k}-i}{j}\int_0^1\sum_{g=j+1}^{i+j}(-1)^{g+1}\binom{i-1}{g-j-1}t^{\mathbf{k}-g}dt=\binom{\mathbf{k}-i}{j}i\int_0^1\left((-1)^j t^{\mathbf{k}-j-1}\left(\frac{t}{t-1}\right)^{1-i}\right)dt=\binom{\mathbf{k}-i}{j}i\frac{(-1)^{j+i+1}\Gamma(i)\Gamma(\mathbf{k}-j-i+1)}{\Gamma(\mathbf{k}-j+1)}=\frac{(-1)^{j+i+1}i!(\mathbf{k}-j-i)!(\mathbf{k}-i)!}{(\mathbf{k}-j)!j!(\mathbf{k}-j-i)!}=(-1)^{j+i+1}\frac{i!(\mathbf{k}-i)!}{\mathbf{k}!}\frac{\mathbf{k}!}{(\mathbf{k}-j)!j!}=\binom{\mathbf{k}}{i}^{-1}(-1)^{1+i}\binom{\mathbf{k}}{j}(-1)^j$.

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

□

The component quasi-distribution, ξ_Δ, is closely related to Ξ_Δ, which is the pairwise difference distribution, since $\sum_{\bar{\Delta}=-\left(\frac{\mathbf{k}}{\frac{3+(-1)^\mathbf{k}}{2}}\right)^{-1}(-\Delta)^\mathbf{k}}^{\frac{1}{\mathbf{k}}(-\Delta)^\mathbf{k}} f_{\xi_\Delta}(\bar{\Delta})=f_{\Xi_\Delta}(\Delta)$. Recall that Theorem .1 established that $f_{\Xi_\Delta}(\Delta)$ is monotonic increasing with a mode at zero if the original distribution is unimodal, $f_{\Xi_{-\Delta}}(-\Delta)$ is thus monotonic decreasing with a mode at zero. In general, if assuming the shape of ξ_Δ is uniform, $\Xi_\mathbf{k}$ is monotonic left and right around zero. The median of $\Xi_\mathbf{k}$ also exhibits a strong tendency to be close to zero, as it can be cast as a weighted mean of the medians of ξ_Δ. When $-\Delta$ is small, all values of ξ_Δ are close to zero, resulting in the median of ξ_Δ being close to zero as well. When $-\Delta$ is large, the median of ξ_Δ depends on its skewness, but the corresponding weight is much smaller, so even if ξ_Δ is highly skewed, the median of $\Xi_\mathbf{k}$ will only be slightly shifted from zero. Denote the median of $\Xi_\mathbf{k}$ as $m\mathbf{k}m$, for the five parametric distributions here, $|m\mathbf{k}m|$s are all $\le 0.1\sigma$ for Ξ_3 and Ξ_4, where σ is the standard deviation of $\Xi_\mathbf{k}$ (SI Dataset S1). Assuming $m\mathbf{k}m=0$, for the even ordinal central moment kernel distribution, the average probability density on the left side of zero is greater than that on the right side, since $\frac{\frac{1}{2}}{\binom{\mathbf{k}}{2}^{-1}(Q(0)-Q(1))^\mathbf{k}}>\frac{\frac{1}{2}}{\frac{1}{\mathbf{k}}(Q(0)-Q(1))^\mathbf{k}}$. This means that, on average, the inequality $f(Q(\epsilon))\ge f(Q(1-\epsilon))$ 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. In all, the monotonic decreasing of the negative pairwise difference distribution guides the general shape of the $\mathbf{k}$th central moment kernel distribution, $\mathbf{k}>2$, forcing it to be unimodal-like with the mode and median close to zero, then, the inequality $f(Q(\epsilon))\le f(Q(1-\epsilon))$ or $f(Q(\epsilon))\ge f(Q(1-\epsilon))$ holds in general. If a distribution is νth γ-ordered and all of its central moment kernel distributions are also νth γ-ordered, it is called completely νth γ-ordered. Although strict complete νth orderliness is difficult to prove, even if the inequality may be violated in a small range, as discussed in Subsection **??**, the mean-SWA$_\epsilon$-median inequality remains valid, in most cases, for the central moment kernel distribution.

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

Proof. Recall that for the $\mathbf{k}$th central moment, the kernel is $\psi_\mathbf{k}(x_1,\ldots,x_\mathbf{k})=\sum_{j=0}^{\mathbf{k}-2}(-1)^j\left(\frac{1}{\mathbf{k}-j}\right)\sum\left(x_{i_1}^{\mathbf{k}-j}x_{i_2}\ldots x_{i_{j+1}}\right)+(-1)^{\mathbf{k}-1}(\mathbf{k}-1)x_1\ldots x_\mathbf{k}$, where the second summation is over $i_1,\ldots,i_{j+1}=1$ to $\mathbf{k}$ with $i_1\ne i_2\ne\ldots\ne i_{j+1}$ and $i_2<i_3<\ldots<i_{j+1}$ (10).

$\psi_\mathbf{k}$ consists of two parts. The first part, $\sum_{j=0}^{\mathbf{k}-2}(-1)^j\left(\frac{1}{\mathbf{k}-j}\right)\sum\left(x_{i_1}^{\mathbf{k}-j}x_{i_2}\ldots x_{i_{j+1}}\right)$, involves a double summation over certain terms. The second part, $(-1)^{\mathbf{k}-1}(\mathbf{k}-1)x_1\ldots x_\mathbf{k}$, carries an alternating sign $(-1)^{\mathbf{k}-1}$ and involves multiplication of the constant $\mathbf{k}-1$ with the product of all the x variables, $x_1x_2\ldots x_\mathbf{k}$. Consider each multiplication cluster $(-1)^j\left(\frac{1}{\mathbf{k}-j}\right)\sum\left(x_{i_1}^{\mathbf{k}-j}x_{i_2}\ldots x_{i_{j+1}}\right)$ for j ranging from 0 to $\mathbf{k}-2$ in the first part. Let each cluster form a single group. The first part can be divided into $\mathbf{k}-1$ groups. Combine this with the second part $(-1)^{\mathbf{k}-1}(\mathbf{k}-1)x_1\ldots x_\mathbf{k}$. Together, the terms of $\psi_\mathbf{k}$ can be divided into a total of $\mathbf{k}$ groups. From the 1st to $\mathbf{k}-1$th group, the gth group has $\binom{\mathbf{k}}{g}\binom{g}{1}$ terms having the form $(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}x_{i_1}^{\mathbf{k}-g+1}x_{i_2}\ldots x_{i_g}$. The final $\mathbf{k}$th group is the term $(-1)^{\mathbf{k}-1}(\mathbf{k}-1)x_1\cdots x_\mathbf{k}$.

There are two ways to divide $\psi_\mathbf{k}$ into $\mathbf{k}$ groups according to the form of each term. The first choice is, if $\mathbf{k}\ne g$, the gth group of $\psi_\mathbf{k}$ has $\binom{\mathbf{k}-l}{g-l}$ terms having the form $(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}x_{i_1}^{\mathbf{k}-g+1}x_{i_2}\cdots x_{i_l}x_{i_{l+1}}\ldots x_{i_g}$, where $x_{i_1},x_{i_2},\cdots,x_{i_l}$ are fixed, $x_{i_{l+1}},\cdots,x_{i_g}$ are selected such that $i_{l+1},\cdots,\ i_g\neq i_1,i_2,\cdots,i_l$ and $i_{l+1}\neq\ldots\neq i_g$. Define another function $\Psi_{\mathbf{k}}\left(x_{i_1},x_{i_2},\cdots,x_{i_l},x_{i_{l+1}},\cdots,x_{i_g}\right)=\left(\lambda x_{i_1}+\mu\right)^{\mathbf{k}-g+1}\left(\lambda x_{i_2}+\mu\right)\cdots\left(\lambda x_{i_l}+\mu\right)\left(\lambda x_{i_{l+1}}+\mu\right)\cdots\left(\lambda x_{i_g}+\mu\right)$, the first group of $\Psi_{\mathbf{k}}$ is $\lambda^{\mathbf{k}}x_{i_1}\cdots x_{i_l}x_{i_{l+1}}\cdots x_{i_g}$, the hth group of $\Psi_{\mathbf{k}}$, $h>1$, has $\binom{\mathbf{k}-g+1}{\mathbf{k}-h-l+2}$ terms having the form $\lambda^{\mathbf{k}-h+1}\mu^{h-1}x_{i_1}^{\mathbf{k}-h-l+2}x_{i_2}\cdots x_{i_l}$. Transforming $\psi_{\mathbf{k}}$ by $\Psi_{\mathbf{k}}$, then combing all terms with $\lambda^{\mathbf{k}-h+1}\mu^{h-1}x_{i_1}^{\mathbf{k}-h-l+2}x_{i_2}\cdots x_{i_l}$, $\mathbf{k}-h-l+2>1$, the summed coefficient is $S1_l=\sum_{g=l}^{h+l-1}(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}\binom{\mathbf{k}-g+1}{\mathbf{k}-h-l+2}\binom{\mathbf{k}-l}{g-l}=\sum_{g=l}^{h+l-1}(-1)^{g+1}\frac{(\mathbf{k}-l)!}{(h+l-g-1)!(\mathbf{k}-h-l+2)!(g-l)!}=0$, since the summation is starting from l, ending at $h+l-1$, the first term includes the factor $g-l=0$, the final term includes the factor $h+l-g-1=0$, the terms in the middle are also zero due to the factorial property.

Another possible choice is the gth group of $\psi_{\mathbf{k}}$ has $(\mathbf{k}-h)\binom{h-1}{g-\mathbf{k}+h-1}$ terms having the form

$$(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}x_{i_1}x_{i_2}\cdots x_{i_j}^{\mathbf{k}-g+1}\cdots x_{i_{\mathbf{k}-h+1}}x_{i_{\mathbf{k}-h+2}}\cdots x_{i_g},$$

provided that $\mathbf{k}\neq g$, $2\leq j\leq\mathbf{k}-h+1$, where $x_{i_1},\ldots,x_{i_{\mathbf{k}-h+1}}$ are fixed, $x_{i_j}^{\mathbf{k}-g+1}$ and $x_{i_{\mathbf{k}-h+2}},\cdots,x_{i_g}$ are selected such that $i_{\mathbf{k}-h+2},\cdots,\ i_g\neq i_1,i_2,\cdots,i_{\mathbf{k}-h+1}$ and $i_{\mathbf{k}-h+2}\neq\ldots\neq i_g$. Transforming these terms by $\Psi_{\mathbf{k}}\left(x_{i_1},x_{i_2},\ldots,x_{i_j},\ldots,x_{i_{\mathbf{k}-h+1}},x_{i_{\mathbf{k}-h+2}},\ldots,x_{i_g}\right)=\left(\lambda x_{i_1}+\mu\right)\left(\lambda x_{i_2}+\mu\right)\cdots\left(\lambda x_{i_j}+\mu\right)^{\mathbf{k}-g+1}\cdots\left(\lambda x_{i_{\mathbf{k}-h+1}}+\mu\right)\left(\lambda x_{i_{\mathbf{k}-h+2}}+\mu\right)\cdots\left(\lambda x_{i_g}+\mu\right)$, then there are $\mathbf{k}-g+1$ terms having the form $\lambda^{\mathbf{k}-h+1}\mu^{h-1}x_{i_1}x_{i_2}\ldots x_{i_{\mathbf{k}-h+1}}$. Transforming the final $\mathbf{k}$th group of $\psi_{\mathbf{k}}$ by $\Psi_{\mathbf{k}}(x_1,\ldots,x_{\mathbf{k}})=(\lambda x_1+\mu)\cdots(\lambda x_{\mathbf{k}}+\mu)$, then, there is one term having the form $(-1)^{\mathbf{k}-1}(\mathbf{k}-1)\lambda^{\mathbf{k}-h+1}\mu^{h-1}x_1x_2\ldots x_{\mathbf{k}-h+1}$. Another possible combination is that the gth group of $\psi_{\mathbf{k}}$ contains $(g-\mathbf{k}+h-1)\binom{h-1}{g-\mathbf{k}+h-1}$ terms having the form $(-1)^{g+1}\frac{1}{\mathbf{k}-g+1}x_{i_1}x_{i_2}\cdots x_{i_{\mathbf{k}-h+1}}x_{i_{\mathbf{k}-h+2}}\cdots x_{i_j}^{\mathbf{k}-g+1}\cdots x_{i_g}$. Transforming these terms by $\Psi_{\mathbf{k}}\left(x_{i_1},x_{i_2},\ldots,x_{i_{\mathbf{k}-h+1}},x_{i_{\mathbf{k}-h+2}},\ldots,x_{i_j},\ldots,x_{i_g}\right)=\left(\lambda x_{i_1}+\mu\right)\left(\lambda x_{i_2}+\mu\right)\cdots\left(\lambda x_{i_{\mathbf{k}-h+1}}+\mu\right)\left(\lambda x_{i_{\mathbf{k}-h+2}}+\mu\right)\cdots\left(\lambda x_{i_j}+\mu\right)^{\mathbf{k}-g+1}\cdots\left(\lambda x_{i_g}+\mu\right)$, then there is only one term having the form $\lambda^{\mathbf{k}-h+1}\mu^{h-1}x_{i_1}x_{i_2}\ldots x_{i_{\mathbf{k}-h+1}}$. The above summation $S1_l$ should also be included, i.e., $x_{i_1}^{\mathbf{k}-h-l+2}=x_{i_1}$, $\mathbf{k}=h+l-1$. So, combing all terms with $\lambda^{\mathbf{k}-h+1}\mu^{h-1}x_{i_1}x_{i_2}\ldots x_{i_{\mathbf{k}-h+1}}$, according to the binomial theorem, the summed coefficient is $S2_l=\sum_{g=\mathbf{k}-h+1}^{\mathbf{k}-1}(-1)^{g+1}\binom{h-1}{g-\mathbf{k}+h-1}\left(\mathbf{k}-h+1+\frac{g-\mathbf{k}+h-1}{\mathbf{k}-g+1}\right)+(-1)^{\mathbf{k}-1}(\mathbf{k}-1)=(\mathbf{k}-h+1)\sum_{g=\mathbf{k}-h+1}^{\mathbf{k}-1}(-1)^{g+1}\binom{h-1}{g-\mathbf{k}+h-1}+\sum_{g=\mathbf{k}-h+1}^{\mathbf{k}-1}(-1)^{g+1}\binom{h-1}{g-\mathbf{k}+h-1}\left(\frac{g-\mathbf{k}+h-1}{\mathbf{k}-g+1}\right)+(-1)^{\mathbf{k}-1}(\mathbf{k}-1)=(-1)^{\mathbf{k}}(\mathbf{k}-h+1)+(h-2)(-1)^{\mathbf{k}}+(-1)^{\mathbf{k}-1}(\mathbf{k}-1)=0$. The summation identities required are $\sum_{g=\mathbf{k}-h+1}^{\mathbf{k}-1}(-1)^{g+1}\binom{h-1}{g-\mathbf{k}+h-1}=(-1)^{\mathbf{k}}$ and $\sum_{g=\mathbf{k}-h+1}^{\mathbf{k}-1}(-1)^{g+1}\binom{h-1}{g-\mathbf{k}+h-1}\left(\frac{g-\mathbf{k}+h-1}{\mathbf{k}-g+1}\right)=(h-2)(-1)^{\mathbf{k}}$. These two summation identities are proven in Lemma **??** and **??**.

Thus, no matter in which way, all terms including μ can be canceled out. The proof is complete by noticing that the remaining part is $\lambda^{\mathbf{k}}\psi_{\mathbf{k}}(x_1,\cdots,x_{\mathbf{k}})$.

□

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

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

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

Congruent Distribution

In the realm of nonparametric statistics, the relative differences, or orders, of robust estimators are of primary importance. A key implication of this principle is that when there is a shift in the parameters of the underlying distribution, all nonparametric estimates should asymptotically change in the same direction, if they are estimating the same attribute of the distribution. If, on the other hand, the mean suggests an increase in the location of the distribution while the median indicates a decrease, a contradiction arises. It is worth noting that such contradiction is not possible for any LL-statistics in a location-scale distribution, as explained in Theorem **??** and **??**. However, it is possible to construct counterexamples to the aforementioned implication in a shape-scale distribution. In the case of the Weibull distribution, its quantile function is $Q_{Wei}(p)=\lambda(-\ln(1-p))^{1/\alpha}$, where $0\leq p\leq1$, $\alpha>0$, $\lambda>0$, λ is a scale parameter, α is a shape parameter, ln is the natural logarithm function. Then, $m=\lambda\sqrt[\alpha]{\ln(2)}$, $\mu=\lambda\Gamma\left(1+\frac{1}{\alpha}\right)$, where Γ is the gamma function. When $\alpha=1$, $m=\lambda\ln(2)\approx0.693\lambda$, $\mu=\lambda$, when $\alpha=\frac{1}{2}$, $m=\lambda\ln^2(2)\approx0.480\lambda$, $\mu=2\lambda$, the mean increases as α changes from 1 to $\frac{1}{2}$, but the median decreases. In the last section, the fundamental role of quantile average was demonstrated by using the method of classifying distributions through the signs of derivatives. To avoid such scenarios, this method can also be used. Let the quantile average function of a parametric distribution be denoted as $\mathrm{QA}(\epsilon,\gamma,\alpha_1,\cdots,\alpha_i,\cdots,\alpha_k)$, where α_i represent the parameters of the distribution, then, a distribution is γ-congruent if and only if the sign of $\frac{\partial\mathrm{QA}}{\partial\alpha_i}$ remains the same for all $0\leq\epsilon\leq\frac{1}{1+\gamma}$. If $\frac{\partial\mathrm{QA}}{\partial\alpha_i}$ is equal to zero or undefined, it can be considered both positive and negative, and thus does not impact the analysis. A distribution is completely γ-congruent if and only if it is γ-congruent and all its central moment kernel distributions are also γ-congruent. Setting $\gamma=1$ constitutes the definitions of congruence and complete congruence. Replacing the QA with γmHLM (defined in the following section) gives the definition of γ-U-congruence. Chebyshev's inequality implies that, for any probability distributions with finite second moments, as the parameters change, even if some LL-statistics change in a direction different from that of the population mean, the magnitude of the changes in

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

Errors	$\bar{x}$	TM	H-L	SM	HM	WM	SQM	BM	MoM	MoRM	mHLM	$rm_{exp,\text{BM}}$	$qm_{exp,\text{BM}}$
WASAB	0.000	0.107	0.088	0.078	0.078	0.066	0.048	0.048	0.034	0.035	0.034	0.002	0.003
WRMSE	0.014	0.111	0.092	0.083	0.083	0.070	0.053	0.053	0.041	0.041	0.038	0.017	0.018
WASB$_{n=5184}$	0.000	0.108	0.089	0.078	0.079	0.066	0.048	0.048	0.034	0.036	0.033	0.002	0.003
WSE $\vee$ WSSE	0.014	0.014	0.014	0.015	0.014	0.014	0.014	0.015	0.017	0.014	0.014	0.017	0.017

Errors		HFM$_\mu$	MP$_\mu$	rm	qm	im	var	var_{bs}	Tsd^2	HFM$_{\mu_2}$	MP$_{\mu_2}$	$rvar$	$qvar$	$ivar$
WASAB		0.037	0.043	0.001	0.002	0.001	0.000	0.000	0.200	0.027	0.042	0.005	0.018	0.003
WRMSE		0.049	0.055	0.015	0.015	0.014	0.017	0.017	0.198	0.042	0.062	0.019	0.026	0.019
WASB$_{n=5184}$		0.038	0.043	0.001	0.002	0.001	0.000	0.001	0.198	0.027	0.043	0.005	0.018	0.003
WSE $\vee$ WSSE		0.018	0.021	0.015	0.015	0.014	0.017	0.017	0.015	0.024	0.032	0.018	0.017	0.018

Errors	tm	tm_{bs}	HFM$_{\mu_3}$	MP$_{\mu_3}$	rtm	qtm	itm	fm	fm_{bs}	HFM$_{\mu_4}$	MP$_{\mu_4}$	rfm	qfm	ifm
WASAB	0.000	0.000	0.052	0.059	0.006	0.083	0.034	0.000	0.000	0.037	0.046	0.024	0.038	0.011
WRMSE	0.019	0.018	0.063	0.074	0.018	0.083	0.044	0.026	0.023	0.049	0.062	0.037	0.043	0.029
WASB$_{n=5184}$	0.001	0.003	0.052	0.059	0.007	0.082	0.038	0.001	0.009	0.037	0.047	0.024	0.036	0.013
WSE $\vee$ WSSE	0.019	0.018	0.021	0.091	0.015	0.012	0.017	0.024	0.021	0.020	0.027	0.021	0.020	0.022

The first table presents the use of the exponential distribution as the consistent distribution for five common unimodal distributions: Weibull, gamma, Pareto, lognormal, and generalized Gaussian distributions. Popular robust mean estimators discussed in REDS 1 were used as comparisons. The breakdown points of mean estimators in the first table, besides H-L estimator and Huber M-estimator, are all $\frac{1}{8}$. The second and third tables present the use of the Weibull distribution as the consistent distribution not plus/plus using the lognormal distribution for the odd ordinal moments optimization and the generalized Gaussian distribution for the even ordinal moments optimization. SQM is the robust mean estimator used in recombined/quantile moments. Unbiased sample central moments (var, tm, fm), U-central moments with quasi-bootstrap (var_{bs}, tm_{bs}, fm_{bs}), and other estimators were used as comparisons. The generalized Gaussian distribution was excluded for He and Fung M-Estimator and Marks percentile estimator, since the logarithmic function does not produce results for negative inputs. The breakdown points of estimators in the second and third table, besides M-estimators and percentile estimator, are all $\frac{1}{24}$. The tables include the average standardized asymptotic bias (ASAB, as $n \to \infty$), root mean square error (RMSE, at $n = 5184$), average standardized bias (ASB, at $n = 5184$) and variance (SE $\vee$ SSE, at $n = 5184$) of these estimators, all reported in the units of the standard deviations of the distribution or corresponding kernel distributions. W means that the results were weighted by the number of Google Scholar search results on May 30, 2022 (including synonyms). The calibrations of d values and the computations of ASAB, ASB, and SSE were described in Subsection , **??** and SI Methods. Detailed results and related codes are available in SI Dataset S1 and GitHub.

the LL-statistics remains bounded compared to the changes in the population mean. Furthermore, distributions with infinite moments can be γ-congruent, since the definition is based on the quantile average, not the population mean.

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

Theorem .4. *A symmetric distribution is always congruent and U-congruent.*

Proof. As shown in Theorem **??** and Theorem **??**, for any symmetric distribution, all quantile averages and all γmHLMs conincide. The conclusion follows immediately. ☐

Theorem .5. *A positive definite location-scale distribution is always γ-congruent.*

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

For the Pareto distribution, $\frac{\partial Q}{\partial \alpha} = \frac{x_m(1-p)^{-1/\alpha}\ln(1-p)}{\alpha^2}$. Since $\ln(1-p) < 0$ for all $0 < p < 1$, $(1-p)^{-1/\alpha} > 0$ for all $0 < p < 1$ and $\alpha > 0$, so $\frac{\partial Q}{\partial \alpha} < 0$, and therefore $\frac{\partial \text{QA}}{\partial \alpha} < 0$, the Pareto distribution is γ-congruent. It is also γ-U-congruent, since γmHLM can also express as a function of $Q(p)$. For the lognormal distribution, $\frac{\partial \text{QA}}{\partial \sigma} = \frac{1}{2}\left(\sqrt{2}\text{erfc}^{-1}(2\gamma\epsilon)\left(-e^{\frac{\sqrt{2}\mu-2\sigma\text{erfc}^{-1}(2\gamma\epsilon)}{\sqrt{2}}}\right) + \left(-\sqrt{2}\right)\text{erfc}^{-1}(2(1-\epsilon))e^{\frac{\sqrt{2}\mu-2\sigma\text{erfc}^{-1}(2(1-\epsilon))}{\sqrt{2}}}\right)$. Since the inverse complementary error function is positive when the input is smaller than 1, and negative when the input is larger than 1, and symmetry around 1, if $0 \le \gamma \le 1$, $\text{erfc}^{-1}(2\gamma\epsilon) \ge -\text{erfc}^{-1}(2-2\epsilon)$, $e^{\mu-\sqrt{2}\sigma\text{erfc}^{-1}(2-2\epsilon)} > e^{\mu-\sqrt{2}\sigma\text{erfc}^{-1}(2\gamma\epsilon)}$. Therefore, if $0 \le \gamma \le 1$, $\frac{\partial \text{QA}}{\partial \sigma} > 0$, the lognormal distribution is γ-congruent. Theorem .4 implies that the generalized Gaussian distribution is congruent and U-congruent. For the Weibull distribution, when α changes from 1 to $\frac{1}{2}$, the average probability density on the left side of the median increases, since $\frac{\frac{1}{2}}{\lambda\ln(2)} < \frac{\frac{1}{2}}{\lambda\ln^2(2)}$, but the mean increases, indicating that the distribution is more heavy-tailed, the probability density of large values will also increase. So, the reason for non-congruence of the Weibull distribution lies in the simultaneous increase of probability densities on two opposite sides as the shape parameter changes: one approaching the bound zero and the other approaching infinity. Note that the gamma distribution does not have this issue, Numerical results indicate that it is likely to be congruent.

The next theorem shows an interesting relation between congruence and the central moment kernel distribution.

Theorem .6. *The second central moment kernal distribution derived from a continuous location-scale unimodal distribution is always γ-congruent.*

Proof. Theorem .3 shows that the central moment kernel distribution generated from a location-scale distribution is also a location-scale distribution. Theorem .1 shows that it is positively definite. Implementing Theorem 12 in REDS 1 yields the desired result. □

Although some parametric distributions are not congruent, as shown in REDS 1. In REDS 1, Theorem 12 establishes that γ-congruence always holds for a positive definite location-scale family distribution and thus for the second central moment kernel distribution generated from a location-scale unimodal distribution as shown in Theorem .6. Theorem .2 demonstrates that all central moment kernel distributions are unimodal-like with mode and median close to zero, as long as they are generated from unimodal distributions. Assuming finite moments and constant $Q(0)-Q(1)$, increasing the mean of a distribution will result in a generally more heavy-tailed distribution, i.e., the probability density of the values close to $Q(1)$ increases, since the total probability density is 1. In the case of the $\mathbf{k}$th central moment kernel distribution, $\mathbf{k} > 2$, while the total probability density on either side of zero remains generally constant as the median is generally close to zero and much less impacted by increasing the mean, the probability density of the values close to zero decreases as the mean increases. This transformation will increase nearly all symmetric weighted averages, in the general sense. Therefore, except for the median, which is assumed to be zero, nearly all symmetric weighted averages for all central moment kernel distributions derived from unimodal distributions should change in the same direction when the parameters change.

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 (7, 23–31). Daniell (1920) stated (24) 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 (7, 30), argued that meaningful comparisons of the efficiencies of various kinds of location estimators can be accomplished by studying their standardized variances, asymptotic variances, and efficiency bounds. Standardized variance, $\frac{\mathrm{Var}(\hat{\theta})}{\theta^2}$, allows the use of simulation studies or empirical data to compare the variances of estimators of distinct parameters. However, a limitation of this approach is the inverse square dependence of the standardized variance on θ. If $\mathrm{Var}\left(\hat{\theta}_1\right) = \mathrm{Var}\left(\hat{\theta}_2\right)$, but θ_1 is close to zero and θ_2 is relatively large, their standardized variances will still differ dramatically. Here, the scaled standard error (SSE) is proposed as a method for estimating the variances of estimators measuring the same attribute, offering a standard error more comparable to that of the sample mean and much less influenced by the magnitude of θ.

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

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

The SSEs of all robust estimators proposed here are often, although many exceptions exist, between those of the sample median and those of the sample mean or median central moments and U-central moments (SI Dataset S1). This is because similar monotonic relations between breakdown point and variance are also very common, e.g., Bickel and Lehmann (7) 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) (32, 33) 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. 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. So, unlike bias, the variance-optimal choice can be very different for distributions with different kurtosis.

Due to combinatorial explosion, the bootstrap (34), introduced by Efron in 1979, is indispensable for computing central moments in practice. In 1981, Bickel and Freedman (35) showed that the bootstrap is asymptotically valid to approximate the original distribution in a wide range of situations, including U-statistics. The limit laws of bootstrapped trimmed U-statistics were proven by Helmers, Janssen, and Veraverbeke (1990) (36). In REDS I, the advantages of quasi-bootstrap were discussed (37–39). By using quasi-sampling, the impact of the number of repetitions of the bootstrap, or bootstrap size, on variance is very small (SI Dataset S1). An estimator based on the quasi-bootstrap approach can be seen as a complex deterministic estimator that is not only computationally efficient but also statistical efficient. The only drawback of quasi-bootstrap compared to non-bootstrap is that a small bootstrap size can produce additional finite sample bias (SI Text).

Discussion

Moments, including raw moments, central moments, and standardized moments, are the most common parameters that describe probability distributions. Central moments are preferred over raw moments because they are invariant to translation. In 1947, Hsu and Robbins proved that the arithmetic mean converges completely to the population mean provided the second moment is finite (40). The strong law of large numbers (proven by Kolmogorov in 1933) (41) implies that the **k**th 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) (42), Pillai and Meng (2016) (43), Cohen, Davis, and Samorodnitsky (2020) (44), and Brown, Cohen, Tang, and Yam (2021) (45). Lindquist and Rachev (2021) raised a critical question in their inspiring comment to Brown et al's paper (45): "What are the proper measures for the location, spread, asymmetry, and dependence (association) for random samples with infinite mean?" (46). From a different perspective, this question closely aligns with the essence of Bickel and Lehmann's open question in 1979 (4). They suggested using median, interquartile range, and medcouple (47) as the robust versions of the first three moments. While answering this question is not the focus of this paper, it is almost certain that the estimators proposed in this series will have a place. Since the efficiency of an L-statistic to the sample mean is generally monotonic with respect to the breakdown point (7), and the estimation of central moments can be transformed into the location estimation of the central moment kernel distribution, similar monotonic relations can be expected. In the case of a distribution with an infinite mean, non-robust estimators will not converge and will not provide valid estimates since their variances will be infinitely large. Therefore, the desired measures should be as robust as possible. Clearly now, if one wants to preserve the original relationship between each moment while ensuring maximum robustness, the natural choices are median, median variance, and median skewness. Similar to the robust version of L-moment (48) being trimmed L-moment (16), mean and central moments now also have their standard most robust version based on the complete congruence of the underlying distribution.

Methods

Data and Software Availability. Data for Table 1 are given in SI Dataset S1-S4. All codes have been deposited in GitHub.

1. FR Hampel, The influence curve and its role in robust estimation. *J. american statistical association* **69**, 383–393 (1974).
2. CF Gauss, Bestimmung der genauigkeit der beobachtungen. *Ibidem* pp. 129–138 (1816).
3. PJ Bickel, EL Lehmann, Descriptive statistics for nonparametric models. iii. dispersion in *Selected works of EL Lehmann.* (Springer), pp. 499–518 (2012).
4. PJ Bickel, EL Lehmann, Descriptive statistics for nonparametric models iv. spread in *Selected Works of EL Lehmann.* (Springer), pp. 519–526 (2012).
5. H Oja, On location, scale, skewness and kurtosis of univariate distributions. *Scand. J. statistics* pp. 154–168 (1981).
6. H Oja, Descriptive statistics for multivariate distributions. *Stat. & Probab. Lett.* **1**, 327–332 (1983).
7. PJ Bickel, EL Lehmann, Descriptive statistics for nonparametric models ii. location in *selected works of EL Lehmann.* (Springer), pp. 473–497 (2012).
8. W van Zwet, Convex transformations: A new approach to skewness and kurtosis in *Selected Works of Willem van Zwet.* (Springer), pp. 3–11 (2012).
9. PJ Rousseeuw, C Croux, Alternatives to the median absolute deviation. *J. Am. Stat. association* **88**, 1273–1283 (1993).
10. PM Heffernan, Unbiased estimation of central moments by using u-statistics. *J. Royal Stat. Soc. Ser. B (Statistical Methodol.* **59**, 861–863 (1997).
11. J Hodges, E Lehmann, Matching in paired comparisons. *The Annals Math. Stat.* **25**, 787–791 (1954).
12. AL Bowley, *Elements of statistics.* (King) No. 8, (1926).
13. WR van Zwet, *Convex Transformations of Random Variables: Nebst Stellingen.* (1964).
14. RA Groeneveld, G Meeden, Measuring skewness and kurtosis. *J. Royal Stat. Soc. Ser. D (The Stat.* **33**, 391–399 (1984).
15. J SAW, Moments of sample moments of censored samples from a normal population. *Biometrika* **45**, 211–221 (1958).
16. EA Elamir, AH Seheult, Trimmed l-moments. *Comput. Stat. & Data Analysis* **43**, 299–314 (2003).
17. RA Fisher, Moments and product moments of sampling distributions. *Proc. Lond. Math. Soc.* **2**, 199–238 (1930).
18. PR Halmos, The theory of unbiased estimation. *The Annals Math. Stat.* **17**, 34–43 (1946).
19. W Hoeffding, A class of statistics with asymptotically normal distribution. *The Annals Math. Stat.* **19**, 293–325 (1948).
20. RJ Serfling, Generalized l-, m-, and r-statistics. *The Annals Stat.* **12**, 76–86 (1984).
21. E Joly, G Lugosi, Robust estimation of u-statistics. *Stoch. Process. their Appl.* **126**, 3760–3773 (2016).
22. P Laforgue, S Clémençon, P Bertail, On medians of (randomized) pairwise means in *International Conference on Machine Learning.* (PMLR), pp. 1272–1281 (2019).
23. S Newcomb, A generalized theory of the combination of observations so as to obtain the best result. *Am. journal Math.* **8**, 343–366 (1886).
24. P Daniell, Observations weighted according to order. *Am. J. Math.* **42**, 222–236 (1920).
25. F Mosteller, On some useful" inefficient" statistics. *The Annals Math. Stat.* **17**, 377–408 (1946).
26. CR Rao, *Advanced statistical methods in biometric research.* (Wiley), (1952).
27. PJ Bickel, Some contributions to the theory of order statistics in *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability.* Vol. 1, pp. 575–591 (1967).
28. 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).
29. L LeCam, On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *The Annals Math. Stat.* **41**, 802–828 (1970).
30. P Bickel, E Lehmann, Descriptive statistics for nonparametric models i. introduction in *Selected Works of EL Lehmann.* (Springer), pp. 465–471 (2012).
31. 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).
32. EL Lehmann, H Scheffé, Completeness, similar regions, and unbiased estimation-part i in *Selected works of EL Lehmann.* (Springer), pp. 233–268 (2011).
33. EL Lehmann, H Scheffé, *Completeness, similar regions, and unbiased estimation—part II.* (Springer), (2012).
34. B Efron, Bootstrap methods: Another look at the jackknife. *The Annals Stat.* **7**, 1–26 (1979).
35. PJ Bickel, DA Freedman, Some asymptotic theory for the bootstrap. *The annals statistics* **9**, 1196–1217 (1981).
36. R Helmers, P Janssen, N Veraverbeke, *Bootstrapping U-quantiles.* (CWI. Department of Operations Research, Statistics, and System Theory [BS]), (1990).
37. RD Richtmyer, A non-random sampling method, based on congruences, for" monte carlo" problems, (New York Univ., New York. Atomic Energy Commission Computing and Applied ...), Technical report (1958).
38. IM Sobol', On the distribution of points in a cube and the approximate evaluation of integrals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki* **7**, 784–802 (1967).
39. KA Do, P Hall, Quasi-random resampling for the bootstrap. *Stat. Comput.* **1**, 13–22 (1991).
40. PL Hsu, H Robbins, Complete convergence and the law of large numbers. *Proc. national academy sciences* **33**, 25–31 (1947).
41. A Kolmogorov, Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.* **4**, 83–91 (1933).
42. M Drton, H Xiao, Wald tests of singular hypotheses. *Bernoulli* **22**, 38–59 (2016).
43. NS Pillai, XL Meng, An unexpected encounter with cauchy and lévy. *The Annals Stat.* **44**, 2089–2097 (2016).
44. 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).
45. 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).
46. WB Lindquist, ST Rachev, Taylor's law and heavy-tailed distributions. *Proc. Natl. Acad. Sci.* **118**, e2118893118 (2021).
47. G Brys, M Hubert, A Struyf, A robust measure of skewness. *J. Comput. Graph. Stat.* **13**, 996–1017 (2004).
48. 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).