

PNAS | June 14, 2023 | vol. XXX | no. XX | 1–3

means, for the Weibull, gamma, Pareto, log-normal and generalized Gaussian distribution,

$$rm_{d=\frac{\mu - \text{WHLM}_{1,k_1,\epsilon_1,\gamma} - \text{WHLM}_{2,k_2,\epsilon_2,\gamma}}{\text{WHLM}_{1,k_1,\epsilon_1,\gamma} - \text{WHLM}_{2,k_2,\epsilon_2,\gamma}}, k_1, k_2, \epsilon_1, \epsilon_2, \gamma, \text{WHLM}_1, \text{WHLM}_2}$$

is consistent for at least one particular case, where  $\mu$ ,  $\text{WHLM}_{1,k_1,\epsilon_1,\gamma}$  and  $\text{WHLM}_{2,k_2,\epsilon_2,\gamma}$  are different location parameters from an exponential distribution. Let  $\text{WHLM}_{1,k_1,\epsilon_1,\gamma} = \text{BM}_{\nu=3,\epsilon=\frac{1}{24}}$ ,

$$\text{WHLM}_{2,k_2,\epsilon_2,\gamma} = m, \text{ then } \mu = \lambda, m = Q\left(\frac{1}{2}\right) = \ln 2\lambda,$$

$$\text{BM}_{\nu=3,\epsilon=\frac{1}{24}} = \lambda \left(1 + \ln \left( \frac{26068394603446272 \sqrt[6]{\frac{\gamma}{247}} \sqrt[3]{11}}{3915/6 \cdot 101898752449325\sqrt{5}} \right) \right),$$

$$\text{the detailed formula is given in the SI Text. So, } d = \frac{\mu - \text{BM}_{\nu=3,\epsilon=\frac{1}{24}}}{\text{BM}_{\nu=3,\epsilon=\frac{1}{24}} - m} = \frac{\lambda - \lambda \left(1 + \ln \left( \frac{26068394603446272 \sqrt[6]{\frac{\gamma}{247}} \sqrt[3]{11}}{3915/6 \cdot 101898752449325\sqrt{5}} \right) \right)}{\lambda \left(1 + \ln \left( \frac{26068394603446272 \sqrt[6]{\frac{\gamma}{247}} \sqrt[3]{11}}{3915/6 \cdot 101898752449325\sqrt{5}} \right) \right) - \ln 2\lambda} =$$

$$-\frac{\ln \left( \frac{26068394603446272 \sqrt[6]{\frac{\gamma}{247}} \sqrt[3]{11}}{3915/6 \cdot 101898752449325\sqrt{5}} \right)}{1 - \ln(2) + \ln \left( \frac{26068394603446272 \sqrt[6]{\frac{\gamma}{247}} \sqrt[3]{11}}{3915/6 \cdot 101898752449325\sqrt{5}} \right)} \approx 0.103. \text{ The biases of}$$

$rm_{d \approx 0.103, \nu=3, \epsilon_1=\frac{1}{24}, \epsilon_2=\frac{1}{24}, \text{BM}, m}$  for distributions with skewness between those of the exponential and symmetric distributions are tiny (SI Dataset S1).  $rm_{d \approx 0.103, \nu=3, \epsilon_1=\frac{1}{24}, \epsilon_2=\frac{1}{24}, \text{BM}, m}$  exhibits excellent performance for all these common unimodal distributions (SI Dataset S1).

The recombined mean is an recombined  $I$ -statistic. Consider an  $I$ -statistic whose LEs are percentiles of a distribution obtained by plugging  $LU$ -statistics into a cumulative distribution function,  $I$  is defined with arithmetic operations, constants and quantile functions, such an estimator is classified as a quantile  $I$ -statistic. One version of the quantile  $I$ -statistic can be defined as  $QI_{d, h_k, k_1, k_2, \epsilon_1, \epsilon_2, \gamma_1, \gamma_2, n, LU_1, LU_2} :=$

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

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

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

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

$WL = \text{WHLM}$ ,  $qm$  is also consistent for any  $\gamma$ -symmetric distributions.

*Proof.* When  $F(WL_{k, \epsilon, \gamma}) \geq \frac{\gamma}{1+\gamma}$ , the solution of  $(F(WL_{k, \epsilon, \gamma}) - \frac{\gamma}{1+\gamma})d + F(WL_{k, \epsilon, \gamma}) = F(\mu)$  is  $d = \frac{F(\mu) - F(WL_{k, \epsilon, \gamma})}{F(WL_{k, \epsilon, \gamma}) - \frac{\gamma}{1+\gamma}}$ . The  $d$  value for the case where  $F(WL_{k, \epsilon, \gamma, n}) < \frac{\gamma}{1+\gamma}$  is the same. The definitions of the location and scale parameters are such that they must satisfy  $F(x; \lambda, \mu) = F(\frac{x-\mu}{\lambda}; 1, 0)$ , then  $F(WL(k, \epsilon, \gamma); \lambda, \mu) = F(\frac{\lambda WL_0(k, \epsilon, \gamma) + \mu - \mu}{\lambda}; 1, 0) = F(WL_0(k, \epsilon, \gamma); 1, 0)$ . It follows that the percentile of any weighted  $L$ -statistic is free of  $\lambda$  and  $\mu$  for a location-scale distribution. Therefore  $d$  in  $qm$  is also invariably a constant. For the  $\gamma$ -symmetric case,  $F(\text{WHLM}_{k, \epsilon, \gamma}) = F(\mu) = F(Q(\frac{\gamma}{1+\gamma})) = \frac{\gamma}{1+\gamma}$  is valid for any  $\gamma$ -symmetric distribution with a finite second moment, as the same values correspond to same percentiles. Then,  $qm_{d, k, \epsilon, \gamma, \text{WHLM}} = F^{-1}((F(\text{WHLM}_{k, \epsilon, \gamma}) - \frac{\gamma}{1+\gamma})d + F(\mu)) = F^{-1}(0 + F(\mu)) = \mu$ . To avoid inconsistency due to post-adjustment,  $F(\mu)$ ,  $F(WL_{k, \epsilon, \gamma})$  and  $\frac{\gamma}{1+\gamma}$  must reside within the range of  $[\gamma\epsilon, 1 - \epsilon]$ . All results are now proven.  $\square$

The cdf of the Pareto distribution is  $F_{Par}(x) = 1 - (\frac{x_m}{x})^\alpha$ . So, set the  $d$  value in  $qm$  with two arbitrary percentiles  $p_1$  and  $p_2$ ,  $d_{Par, qm} =$

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

$\frac{1 - (\frac{\alpha-1}{\alpha})^\alpha - p_1}{p_1 - p_2}$ . The  $d$  value in  $qm$  for the exponential distribution is always identical to  $d_{Par, qm}$  as  $\alpha \rightarrow \infty$ , since  $\lim_{\alpha \rightarrow \infty} (\frac{\alpha-1}{\alpha})^\alpha = \frac{1}{e}$  and the cdf of the exponential distribution is  $F_{exp}(x) = 1 - e^{-\lambda^{-1}x}$ , then  $d_{exp, qm} =$

$$\frac{(1-e^{-1}) - \left( 1 - e^{-\ln(\frac{1}{1-p_1})} \right)}{\left( 1 - e^{-\ln(\frac{1}{1-p_1})} \right) - \left( 1 - e^{-\ln(\frac{1}{1-p_2})} \right)} = \frac{1 - \frac{1}{e} - p_1}{p_1 - p_2}. \text{ So, for the}$$

Weibull, gamma, Pareto, lognormal and generalized Gaussian distribution,  $qm_{d=\frac{F_{exp}(\mu) - F_{exp}(\text{WHLM}_{k, \epsilon, \gamma})}{F_{exp}(\text{WHLM}_{k, \epsilon, \gamma}) - \frac{\gamma}{1+\gamma}}, k, \epsilon, \gamma, \text{WHLM}}$

is also consistent for at least one particular case, provided that  $\mu$  and  $\text{WHLM}_{k, \epsilon, \gamma}$  are different location parameters from an exponential distribution and  $F(\mu)$ ,  $F(\text{WHLM}_{k, \epsilon, \gamma})$  and  $\frac{\gamma}{1+\gamma}$  are all within the range of  $[\gamma\epsilon, 1 - \epsilon]$ . Also let  $\text{WHLM}_{k, \epsilon, \gamma} = \text{BM}_{\nu=3, \epsilon=\frac{1}{24}}$

$$\text{and } \mu = \lambda, \text{ then } d = \frac{F_{exp}(\mu) - F_{exp}(\text{BM}_{\nu=3, \epsilon=\frac{1}{24}})}{F_{exp}(\text{BM}_{\nu=3, \epsilon=\frac{1}{24}}) - \frac{\gamma}{1+\gamma}} =$$

$$\frac{-e^{-1} + e - \left( 1 + \ln \left( \frac{26068394603446272 \sqrt[6]{\frac{\gamma}{247}} \sqrt[3]{11}}{3915/6 \cdot 101898752449325\sqrt{5}} \right) \right)}{-\frac{1}{2} - e - \left( 1 + \ln \left( \frac{26068394603446272 \sqrt[6]{\frac{\gamma}{247}} \sqrt[3]{11}}{3915/6 \cdot 101898752449325\sqrt{5}} \right) \right)} =$$

$$\frac{\frac{101898752449325\sqrt{5} \sqrt[6]{\frac{\gamma}{247}} \sqrt[3]{11}}{26068394603446272} - \frac{1}{e}}{\frac{1}{2} - \frac{101898752449325\sqrt{5} \sqrt[6]{\frac{\gamma}{247}} \sqrt[3]{11}}{26068394603446272} - \frac{1}{e}} \approx 0.088. \quad F_{exp}(\mu),$$

$F_{exp}(\text{BM}_{\nu=3, \epsilon=\frac{1}{24}})$  and  $\frac{1}{2}$  are all within the range of  $[\frac{1}{24}, \frac{23}{24}]$ .  $qm_{d \approx 0.088, \nu=3, \epsilon=\frac{1}{24}, \text{BM}}$  works better in the fat-tail scenarios (SI Dataset S1). Theorem A.1 and A.2 show that  $rm_{d \approx 0.103, \nu=3, \epsilon_1=\frac{1}{24}, \epsilon_2=\frac{1}{24}, \text{BM}, m}$  and  $qm_{d \approx 0.088, \nu=3, \epsilon=\frac{1}{24}, \text{BM}}$

are both consistent mean estimators for any symmetric distribution and the exponential distribution with finite second moments. It's obvious that the asymptotic breakdown points of  $rm_{d \approx 0.103, \nu=3, \epsilon_1=\frac{1}{24}, \epsilon_2=\frac{1}{2}, BM, m}$  and  $qm_{d \approx 0.088, \nu=3, \epsilon=\frac{1}{24}, BM}$  are both  $\frac{1}{24}$ . Therefore they are all invariant means.

To study the impact of the choice of WLS in  $rm$  and  $qm$ , it is constructive to recall that a weighted  $L$ -statistic is a combination of order statistics. While using a less-biased weighted  $L$ -statistic can generally enhance performance (SI Dataset S1), there is a greater risk of violation in the semiparametric framework. However, the mean- $WA_{\epsilon, \gamma}$ - $\gamma$ -median inequality is robust to slight fluctuations of the QA function of the underlying distribution. Suppose for a right-skewed distribution, the QA function is generally decreasing with respect to  $\epsilon$  in  $[0, u]$ , but increasing in  $[u, \frac{1}{1+\gamma}]$ , since all quantile averages with breakdown points from  $\epsilon$  to  $\frac{1}{1+\gamma}$  will be included in the computation of  $WA_{\epsilon, \gamma}$ , as long as  $\frac{1}{1+\gamma} - u \ll \frac{1}{1+\gamma} - \gamma\epsilon$ , and other portions of the QA function satisfy the inequality constraints that define the  $\nu$ th  $\gamma$ -orderliness on which the  $WA_{\epsilon, \gamma}$  is based, if  $0 \leq \gamma \leq 1$ , the mean- $WA_{\epsilon, \gamma}$ - $\gamma$ -median inequality still holds. This is due to the violation of  $\nu$ th  $\gamma$ -orderliness being bounded, when  $0 \leq \gamma \leq 1$ , as shown in RMSM I and therefore cannot be extreme for unimodal distributions with finite second moments. For instance, the SQA function of the Weibull distribution is non-monotonic with respect to  $\epsilon$  when the shape parameter  $\alpha > \frac{1}{1-\ln(2)} \approx 3.259$  as shown in the SI Text of RMSM I, the violation of the second and third orderliness starts near this parameter as well, yet the mean- $BM_{\nu=3, \epsilon=\frac{1}{24}}$ -median inequality retains valid when  $\alpha \leq 3.387$ . Another key factor in determining the risk of violation is the skewness of the distribution. In RSM I, it was demonstrated that in a family of distributions differing by a skewness-increasing transformation in van Zwet's sense, the violation of orderliness, if it happens, often only occurs as the distribution nears symmetry (1).

1. WR van Zwet, *Convex Transformations of Random Variables: Nebst Stellingen*. (1964).