Semiparametric robust mean estimations based on the orderliness of quantile averages

Tuban Lee

This manuscript was compiled on June 10, 2023

semiparametric | mean-median-mode inequality | asymptotic | unimodal | Hodges–Lehmann estimator

Hodges–Lehmann inequality and γ-U-orderliness

The Hodges–Lehmann estimator stands out as a unique robust location estimator due to its definition being substantially dissimilar from conventional L-estimators, R-estimators, and M-estimators. In their landmark paper, *Estimates of location based on rank tests*, Hodges and Lehmann (1) proposed two methods for computing the H-L estimator: the Wilcoxon score R-estimator and the median of pairwise means. The Wilcoxon score R-estimator is a location estimator based on signed-rank test, or R-estimator, (1) and was later independently discovered by Sen (1963) (2, 3). However, the median of pairwise means is a generalized L-statistic and a trimmed U-statistic, as classified by Serfling in his novel conceptualized study in 1984 (4). Serfling further advanced the understanding by generalizing the H-L kernel as $hl_k\left(x_1, \ldots, x_k\right) = \frac{1}{k}\sum_{i=1}^{k} x_i$, where $k \in \mathbb{N}$ (4). Here, the weighted H-L kernel is defined as $whl_k\left(x_1, \ldots, x_k\right) = \frac{\sum_{i=1}^{k} x_i \mathbf{w}_i}{\sum_{i=1}^{k} \mathbf{w}_i}$, where $\mathbf{w}_i$s are the weights applied to each element.

By using the weighted H-L kernel and the L-estimator, it is now clear that the Hodges-Lehmann estimator is an LL-statistic, the definition of which is provided as follows:

$$LL_{k,\epsilon,\gamma,n} \coloneqq L_{\epsilon_0,\gamma,n}\left(\operatorname{sort}\left(\left(whl_k\left(X_{N_1}, \cdots, X_{N_k}\right)\right)_{N=1}^{\binom{n}{k}}\right)\right),$$

where $L_{\epsilon_0,\gamma,n}(Y)$ represents the ϵ_0,γ-L-estimator that uses the sorted sequence, $\operatorname{sort}\left(whl_k\left(X_{N_1}, \cdots, X_{N_k}\right)\right)_{N=1}^{\binom{n}{k}}$, as input. The upper asymptotic breakdown point of $LL_{k,\epsilon,\gamma}$ is $\epsilon = 1-(1-\epsilon_0)^{\frac{1}{k}}$, as proven in DSSM II. There are two ways to adjust the breakdown point: either by setting k as a constant and adjusting ϵ_0, or by setting ϵ_0 as a constant and adjusting k. In the above definition, k is discrete, but the bootstrap method can be applied to ensure the continuity of k, also making the breakdown point continuous. Specifically, if $k \in \mathbb{R}$, let the bootstrap size be denoted by b, then first sampling the original sample $(1-k+\lfloor k \rfloor)b$ times with each sample size of $\lfloor k \rfloor$, and then subsequently sampling $(1-\lceil k \rceil + k)b$ times with each sample size of $\lceil k \rceil$, $(1-k+\lfloor k \rfloor)b \in \mathbb{N}$, $(1-\lceil k \rceil + k)b \in \mathbb{N}$. The corresponding kernels are computed separately, and the pooled sorted sequence is used as the input for the L-estimator.

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

ACKNOWLEDGMENTS. I sincerely acknowledge the insightful comments from the editor which considerably elevated the lucidity and merit of this paper.

1. J Hodges Jr, E Lehmann, Estimates of location based on rank tests. *The Annals Math. Stat.* **34**, 598–611 (1963).
2. PK Sen, On the estimation of relative potency in dilution (-direct) assays by distribution-free methods. *Biometrics* pp. 532–552 (1963).
3. M Ghosh, MJ Schell, PK Sen, A conversation with pranab kumar sen. *Stat. Sci.* pp. 548–564 (2008).
4. RJ Serfling, Generalized l-, m-, and r-statistics. *The Annals Stat.* **12**, 76–86 (1984).

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