Near-consistent robust estimations of moments for unimodal distributions

Tuban Lee

This manuscript was compiled on May 19, 2023

Descriptive statistics for parametric models currently heavily rely on the accuracy of distributional assumptions. Here, leveraging 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 potential inconsistencies between the sample mean ($\bar{x}$) and robust location estimators with non-zero asymptotic breakdown points in distributions with finite moments 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 interesting approach is based on L-estimators, such as percentile estimators. Examples of percentile estimators for the Weibull distribution, the reader is referred to Menon (1963) (6), Dubey (1967) (7), Marks (2005) (8), and Boudt, Caliskan, and Croux (2011) (9)'s works. At the outset of the study of percentile estimators, it was known that they arithmetically utilize the invariant structures of probability distributions (6, 7). Maybe such estimators can be named as I-statistics. Formally, an estimator is classified as an I-statistic if it asymptotically satisfies $\mathrm{I}\left(\mathrm{LE}_1, \ldots, \mathrm{LE}_l\right) = \left(\theta_1, \ldots, \theta_q\right)$ for the distribution it is consistent, where LEs are calculated with the use of LU-statistics (defined in Subsection **??**), I is defined using arithmetic operations and constants but may also incorporate transcendental functions and quantile functions, and θs are the population parameters it estimates. A subclass of I-statistics, arithmetic I-statistics, is defined as LEs are LU-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-estimators, 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 LU-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-estimator 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 for central moments, rendering the approach ill-suited (SI Dataset S1).

The majority of robust location estimators commonly used are symmetric owing to the prevalence of symmetric distributions. An asymmetric weighted L-statistic can achieve consistency for a semiparametric class of skewed distributions; but the lack of symmetry makes it suitable only for certain applications. Shifting from semiparametrics to parametrics, consider an estimator with a non-zero asymptotic breakdown point that is consistent simultanously for both a semiparametric class of distributions and a distinct parametric distribution with finite moments, such a robust location estimator is called an invariant mean. Based on the mean-weighted L-statistic-γ-median inequality, the recombined mean is defined as

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

where d is the key factor for bias correction, γm_n is the sample γ-median, $\mathrm{WL}_{\epsilon,\gamma,n}$ is the weighted L-statistic. If γ is omitted, $\gamma = 1$ is assumed. The subsequent theorem shows the significance of this arithmetic I-statistic.

Significance Statement

Bias, variance, and contamination are the three main errors in statistics. Consistent robust estimation is unattainable without parametric assumptions. In this article, invariant moments are proposed as a means of achieving near-consistent and robust estimations of moments, even in scenarios where moderate violations of distributional assumptions occur, while the variances are sometimes smaller than those of the sample moments.

T.L. designed research, performed research, analyzed data, and wrote the paper.

The author declares no competing interest.

[1]To whom correspondence should be addressed. E-mail: tl@biomathematics.org

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

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

1. CF Gauss, *Theoria combinationis observationum erroribus minimis obnoxiae*. (Henricus Dieterich), (1823).
2. S Newcomb, A generalized theory of the combination of observations so as to obtain the best result. *Am. journal Math.* **8**, 343–366 (1886).
3. S Newcomb, Researches on the motion of the moon. part ii, the mean motion of the moon and other astronomical elements derived from observations of eclipses and occultations extending from the period of the babylonians until ad 1908. *United States. Naut. Alm. Off. Astron. paper; v. 9* **9**, 1 (1912).
4. PJ Huber, Robust estimation of a location parameter. *Ann. Math. Stat.* **35**, 73–101 (1964).
5. X He, WK Fung, Method of medians for lifetime data with weibull models. *Stat. medicine* **18**, 1993–2009 (1999).
6. M Menon, Estimation of the shape and scale parameters of the weibull distribution. *Technometrics* **5**, 175–182 (1963).
7. SD Dubey, Some percentile estimators for weibull parameters. *Technometrics* **9**, 119–129 (1967).
8. NB Marks, Estimation of weibull parameters from common percentiles. *J. applied Stat.* **32**, 17–24 (2005).
9. K Boudt, D Caliskan, C Croux, Robust explicit estimators of weibull parameters. *Metrika* **73**, 187–209 (2011).