

Analysis of Existence and Validity of Unbiased Estimation

Junting Liu

*School of Mathematics and Statistics, Beijing Institute of Technology, Beijing, China
1120223737@bit.edu.cn*

Abstract. Unbiased estimation, as one of the core concepts of parameter estimation in statistics, requires that the mathematical expectation of the estimator be strictly equal to the true value of the estimated parameter. This characteristic gives it the important advantage of "no systematic deviation" in theory. However, in the actual process of statistical inference, the existence, uniqueness and validity of unbiased estimation often face many theoretical limitations and practical challenges. This study systematically sorts out the theoretical basis of unbiased estimation and comprehensively explores the applicable conditions, essential limitations and possible improvement directions of unbiased estimation. The study adopts a method combining theoretical derivation and empirical analysis, focusing on the following key issues: Firstly, under what model settings may unbiased estimation generally exist; Secondly, when unbiased estimation exists, how to evaluate its statistical efficiency and what kind of unbiased estimation should be adopted; Finally, in complex scenarios where traditional unbiased estimation performs poorly, what are the feasible improvement strategies? Through this series of studies, this study aims to provide theoretical basis and practical guidance for statistical practitioners to select appropriate estimation methods in different application scenarios.

Keywords: Existence and Validity, Cramér-Rao Lower Bound, Independent and Identically Distributed Assumption, Large Sample Assumption, Bias Correction.

1. Introduction

As a central part of parameter estimation, unbiased estimation, which is one of the fundamentals of statistics, is widely applied in both theoretical research and practical usage. The definition of unbiased estimation states that the average value of the estimator should be equal to the real parameter. Thus, unbiased estimation has the corresponding theoretical advantage for the lack of bias of systematic bias. Because it can guarantee the absence of systematic errors in long-term statistical inference processes, unbiased estimation has been widely applied in statistics and has become a core component of many classical statistical methods. In statistical analyses across many fields, unbiased estimation provides the theoretical foundation and ensures the correctness and reliability of estimation results. This advantage of unbiased estimators makes them suitable for any kind of statistical analysis. However, as a theoretically perfect attribute, unbiased estimation does have its limitations under certain circumstances in practice--if the sample size is small, the variance

will be high, and estimates tend to lack stability and precision; therefore, how to ensure unbiasedness and effective accuracy with minimum variance remains a difficult problem to solve.

The main issue related to unbiased estimation is that it works or does not work in the sense of whether one gets a reasonable answer. In practice, the quality of an estimator not only relies on whether it is unbiased but how much variation there is around this average. If an unbiased estimator has too high a variance, then even though its average prediction matches the real parameter, it is still producing noisy predictions which do not provide a good representation of the real parameters. For instance, when the sample size is small, if the estimation is unbiased, there can be big variance in the estimation result. In that case, the biased estimation method can produce lower variance, even though it sacrifices unbiasedness, which benefits estimation stability. How to balance unbiasedness and effectiveness both theoretically and practically is currently a topic of great concern to statisticians. On one hand, how to evaluate the statistical efficiency of unbiased estimation; on the other hand, how to choose the appropriate unbiased estimation method according to specific situation.

This study aims to systematically investigate the existence, uniqueness, and effectiveness of unbiased estimation, analyze its performance in different statistical models, and propose corresponding improvement measures. First, this study will review the mathematical definition and basic properties of unbiased estimation, clarifying its core requirements and theoretical advantages. By combining theoretical derivation with empirical analysis, this paper contributes to further development of statistical theory from a new viewpoint and provides both theoretical basis and methodological guidance on selecting suitable methods of estimation when applied to practice. Along with a quick developing environment in big data and machine learning, traditional unbiased estimators encounter numerous unprecedented problems. Thus, this paper attempts to find some feasible approaches to cope with these newly emerged problems about unbiasedness. Finally, this paper attempts to present some innovative findings concerning unbiased estimation in the coming big data and high-dimensional environments to help the related researchers to advance statistics.

2. Literature review

2.1. Unbiased estimation: concept and mathematical definition

An unbiased estimator refers to an estimator whose expected value equals the true value of the parameter being estimated. More formally, let θ be the parameter to be estimated and $\hat{\theta}$ be the estimator derived from sample data. If the expectation of $\hat{\theta}$ equals θ [1].

Mathematically, the definition of an unbiased estimator can be expressed as follows: Let $f(x;\theta)$ denote the probability distribution function of the population, where x is the variable and $\theta \in \Theta$ is the unknown parameter, with Θ referred to as the parameter space. Let $(X_1, X_2, \dots, X_n)$ be a random sample drawn from the population. If $\hat{\theta}(X_1, X_2, \dots, X_n)$ is an estimator of the parameter θ , and for all $\theta \in \Theta$, the equation:

$$E[\hat{\theta}(X_1, X_2, \dots, X_n)] = \theta \quad (1)$$

holds, where $E\left[\hat{\theta}(X_1, X_2, \dots, X_n)\right]$ denotes the mathematical expectation, then $\hat{\theta}$ is called an unbiased estimator of θ , and $\hat{\theta}$ is said to possess unbiasedness. (In the subsequent theoretical derivations of this study, this assumption will be maintained.)

2.2. Biased estimation and its correction

Unlike unbiased estimation, if the expected value of an estimator does not equal the true parameter, then the estimator is referred to as a biased estimator. Simply put, the long-term average of a biased estimator deviates from the true parameter.

For example, the expected value of the sample variance σ_y^2 computed from all possible samples of size n in a population tends to underestimate the true population variance σ^2 [2]. In fact:

$$E\left(\sigma_y^2\right) = E\left[\frac{1}{n} \sum_{i=1}^n \left(y_i - \frac{1}{n} \sum_{j=1}^n y_j\right)^2\right] = \frac{n-1}{n} \sigma^2 \quad (2)$$

Through Bessel's correction, the coefficient is adjusted to eliminate bias [3]. That is:

$$s^2 = \frac{n}{n-1} \sigma_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \quad (3)$$

Then it turns to be an unbiased estimator:

$$E(s^2) = \sigma^2 \quad (4)$$

The essence of this correction lies in adjusting the denominator (degrees of freedom) to compensate for the loss of information caused by using the sample mean (instead of the true population mean). In a more general linear modeling framework, this concept of degrees-of-freedom adjustment extends to the notion of residual degrees of freedom.

2.3. Challenges of unbiased estimation

Traditional unbiased estimators often rely on the classical independent and identically distributed (i.i.d.) assumption, which ensures that each sample has the same distribution and that samples are mutually independent. However, in real-world data, this assumption does not always hold. In many practical problems, data may exhibit autocorrelation, heteroskedasticity, or other complex structural characteristics, posing significant challenges to classical unbiased estimation methods. For example, time-series data in financial markets often exhibit autocorrelation, while data in biology and medical fields may show heteroskedasticity. In these cases, traditional unbiased estimation methods may fail to effectively estimate the true parameters, and their estimation results may have large bias or variance. Therefore, in these complex scenarios, how to improve traditional unbiased estimation methods to enhance their stability and reliability in practical applications has become an urgent problem to solve in statistics.

Moreover, in some complex model settings, the performance of unbiased estimation may not be ideal, especially when the data distribution does not conform to classical assumptions. For example, for high-dimensional data, non-normal distributions, or cases where the model specification is not entirely correct, traditional unbiased estimation may fail to provide ideal estimation results. Thus,

how to improve the performance of unbiased estimation in these complex scenarios and propose new estimation methods has become a major challenge in current statistical inference.

3. The existence of unbiased estimation

In fact, constructing unbiased estimators based solely on definitional requirements presents significant theoretical challenges. However, unbiased estimators are commonly encountered in practical applications because real-world sample spaces typically satisfy either the independent and identically distributed assumption or large-sample properties. Under these conditions, the existence of unbiased estimators becomes generally guaranteed. This study will provide specific construction methodologies.

3.1. Independent and identically distributed assumption

In statistical inference, the independent and identically distributed assumption serves as one of the fundamental premises for constructing unbiased estimators. Suppose to draw n samples $X_1, X_2, \dots, X_n$ from a population, where these samples are mutually independent and share the same distribution, denoted as

$$X_1, X_2, \dots, X_n \sim f(x; \theta) \quad (5)$$

This assumption ensures the elimination of statistical dependencies among sample data, thereby simplifying and improving the accuracy of expectation calculations.

If the data are independent and identically distributed, then each sample X_i shares the same probability density function $f(x_i; \theta)$ and maintains mutual independence. Under these conditions, the expectation of the estimator $\hat{\theta} = g(X_1, X_2, \dots, X_n)$ can be expressed as:

$$E[\hat{\theta}] = E[g(X_1, X_2, \dots, X_n)] = \int_{\mathcal{X}^n} g(x_1, x_2, \dots, x_n) f^n(x_i; \theta) dx_1 \dots dx_n \quad (6)$$

Independence ensures that the joint density function can be factored into the product of individual sample density functions, while identical distribution guarantees that each sample's density function $f(x_i, \theta)$ remains consistent across the entire sample. This approach makes the expectation calculation both feasible and precise.

Moreover, the estimator's design must ensure that its expected value remains accurate across all possible sample configurations, with the unbiasedness requirement mandating that the expectation equals the true parameter value θ :

$$E[\hat{\theta}] = \int_{\mathcal{X}^n} g(x_1, x_2, \dots, x_n) f^n(x_i; \theta) dx_1 \dots dx_n = \theta \quad (7)$$

The i.i.d. assumption enables this objective to be achieved through the aforementioned integral derivation and provides the mathematical foundation for constructing unbiased estimators.

3.2. Asymptotic unbiasedness and the central limit theorem

For finite-sample unbiasedness, the i.i.d. property ensures that the estimator's expected value equals the true parameter θ in any sample of fixed size n . Asymptotic unbiasedness extends this guarantee to large samples: even if an estimator is biased for finite n , its bias may vanish as:

$$n \rightarrow \infty \quad (8)$$

To ensure unbiasedness in large samples, consider an estimator $\hat{\theta}_n$ based on a sample of size n , with expected value $E[\hat{\theta}_n]$. Asymptotic unbiasedness requires:

$$\lim_{n \rightarrow \infty} E[\hat{\theta}_n] = \theta \quad (9)$$

The Central Limit Theorem (CLT) plays a pivotal role in this context. Assuming the estimated parameter is a sample mean or similar statistic, the CLT demonstrates that as the sample size increases, the distribution of the sample mean will approach a normal distribution, with its meaning converging to the true parameter θ while its standard deviation decreases with growing sample size [4].

Specifically, for independent and identically distributed samples $X_1, X_2, \dots, X_n$ and their sample meaning

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n X_i. \quad (10)$$

According to the Central Limit Theorem, when the sample size n is sufficiently large, the distribution of the sample mean $\hat{\theta}_n$ will approximate a normal distribution:

$$\hat{\theta}_n \sim N\left(\theta, \frac{\sigma^2}{n}\right) \quad (11)$$

where σ^2 is the population variance with expected value θ [5]. Consequently, as the sample size n increases, the expected value of the sample means converges to the true parameter value θ , while the standard deviation of the error term progressively decreases. Specifically, the bias of the sample means vanishes asymptotically, satisfying:

$$\hat{\theta}_n - \theta \sim N\left(0, \frac{\sigma^2}{n}\right) \quad (12)$$

This implies that as n increases, the bias tends toward zero, which means:

$$E[\hat{\theta}_n] = \theta \quad (13)$$

Therefore, the Central Limit Theorem provides critical theoretical justification for deriving asymptotic unbiasedness: it guarantees that sample means, or other estimators will approach

unbiased estimation in large sample scenarios, meaning their expected values converge to the true parameter value.

4. Example of unbiased estimation

Consider the estimation of the parameter in a Bernoulli distribution. Let $X_1, X_2, \dots, X_n$ be independent and identically distributed Bernoulli random variables, where each X_i follows a Bernoulli distribution with parameter p . As previously mentioned, the sample mean serves as an unbiased estimator of p in large sample spaces. Now, the sample mean (given by the formula below) is examined using to estimate p in sample spaces of arbitrary size:

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i. \quad (14)$$

Based on the properties of Bernoulli distribution, the expectation of a single sample X_i is $E[X_i] = p$. Consequently, the expectation of the sample means:

$$E[\hat{p}] = E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n E[X_i] = \frac{1}{n} \cdot n \cdot p = p. \quad (15)$$

This demonstrates that the sample mean is unbiased, as its expected value equals the true success probability p , thereby satisfying the unbiasedness condition.

Furthermore, considering the variance of the sample means:

$$\text{Var}(\hat{p}) = \frac{p(1-p)}{n}. \quad (16)$$

As the sample size n increases, the variance of the sample proportion progressively decreases. Simultaneously, this decreasing variance demonstrates that the sample proportion is not only unbiased but also achieves greater estimation precision with larger datasets.

The sample mean, as an unbiased estimator of parameters, holds significant practical importance. It is widely applied in statistical inference, such as constructing confidence intervals and conducting hypothesis tests. Through sample proportions, it enables accurate estimation of the probability of an event occurring in the population.

For instance: In market research, sample proportions can estimate consumer acceptance of a product. In clinical medical trials, sample proportions help assess the cure rate of treatment. In quality control, sample proportions estimate the pass rate in production processes [7]. This demonstrates the sample mean's versatility as a fundamental estimation tool across diverse application domains.

5. The validity of unbiased estimation

As is mentioned in the example, the effectiveness of an estimator depends not merely on its unbiasedness but also on its variance. While an unbiased estimator's expected value equals the true parameter value, its estimation precision which is evaluated by variance is equally crucial. If an estimator exhibits large variance, even when unbiased, it may produce unstable estimates that poorly

reflect the true parameter. Therefore, evaluating the effectiveness of unbiased estimators necessarily requires joint consideration of variance analysis.

The variance, under certain conditions, is bounded by the Cramér-Rao lower bound. The Cramér-Rao inequality establishes a lower bound for the variance of unbiased estimators [8]. Specifically, suppose there exists an unbiased estimator for parameter θ , and the parameter space satisfies certain conditions [9]. The conditions are as follows:

- ① Θ is an open interval on the real line.
- ② The family of distributions has a common support, meaning the set of samples where the probability density function is positive does not depend on θ .
- ③ For all samples, $\forall \theta \in \Theta$, the partial derivative $\frac{\partial f}{\partial \theta}$ exists.
- ④ In the continuous case, the order of integration and partial differentiation can be interchanged for the probability density function $f(x; \theta)$, i.e.:

$$\frac{\partial}{\partial \theta} \int_{\mathbb{R}} f(x; \theta) dx = \int_{\mathbb{R}} \frac{\partial}{\partial \theta} f(x; \theta) dx. \quad (17)$$

In the discrete case, the order of infinite summation and partial differentiation can be interchanged for the probability mass function, i.e.:

$$\frac{\partial}{\partial \theta} \sum f(x; \theta) dx = \sum \frac{\partial}{\partial \theta} f(x; \theta) dx. \quad (18)$$

- ⑤ The second-order moment of the partial derivative of the logarithm of the probability distribution exists, i.e.:

$$E\left[\frac{\partial^2}{\partial \theta^2} \log F(x; \theta)\right] < +\infty. \quad (19)$$

Under these conditions, the Cramér-Rao inequality can be expressed as:

$$\text{Var}(\hat{\theta}) \geq \frac{1}{nI(\theta)}, \quad (20)$$

where $I(\theta)$ is the Fisher information (Lehmann), specifically expressed as:

$$I(\theta) = -E\left[\frac{\partial^2}{\partial \theta^2} \log F(x; \theta)\right] \quad (21)$$

According to this inequality, for any unbiased estimator, the variance cannot be lower than $\frac{1}{nI(\theta)}$, indicating that there exists a theoretical lower bound for the variance of unbiased estimators. The Cramér-Rao lower bound measures the effectiveness of unbiased estimation. It is worth noting that under the same probability distribution function, as n decreases, the variance lower bound increases accordingly. In small sample spaces, the effectiveness of unbiased estimation diminishes.

6. Limitations of unbiased estimation and analysis of improvement measures

As previously mentioned, in small sample spaces, unbiased estimation may exhibit relatively large variance, whereas biased estimation can yield estimates with smaller variance and better stability.

Thus, when the mean value is not considered, biased estimation may demonstrate superior estimation performance.

To address the limitations of unbiased estimation, this study proposes two improvement measures and alternative approaches.

6.1. Minimum Variance Unbiased Estimator (MVUE)

By utilizing the concepts of sufficient statistics and completeness, the minimum variance unbiased estimator can be identified. The Lehmann-Scheffé theorem states that if an unbiased estimator is both sufficient and complete, it will be the unique minimum variance unbiased estimator [10]. This ensures that, while maintaining unbiasedness, the estimator with the smallest variance can be selected.

6.2. Bias correction for biased estimation

Although the expected value of a biased estimator deviates from the true parameter, its bias can be reduced by introducing correction terms [11]. For example, bias correction methods (such as weighted estimation) can incorporate additional steps to mitigate bias and enhance estimation accuracy.

7. Conclusion

This study systematically investigates the theoretical foundations, practical applications, and improvement measures of unbiased estimation, while also identifying certain limitations and areas requiring further research.

From a methodological standpoint, while this study's analytical framework is grounded in conventional statistical inference approaches and offers a thorough examination of established unbiased estimation principles, its coverage of cutting-edge domains like high-dimensional statistical analysis and distributed computational methods within big data contexts shows notable limitations in terms of both scope and profundity. This is particularly evident when dealing with extremely high-dimensional datasets, where fundamental aspects of conventional unbiased estimation techniques - including their theoretical foundations and computational behaviors - demand more rigorous and extensive investigation.

From an innovative research design perspective, although the current study creatively proposes approaches such as minimum variance unbiased estimation and examines several bias corrections approaches to correct biased estimation, there is not enough integration between modern computational methods (such as deep learning, reinforcement learning), and classical unbiased estimation theory. This deficiency of methodological means has certain restrictions on applying these research outcomes to leading fields like AI.

From a methodological standpoint, although this study adopts a research paradigm combining theoretical derivation with numerical simulation and has achieved certain theoretical breakthroughs it lacks empirical validation based on large-scale real-world datasets. Especially in big data scenarios with exponentially growing sample sizes, practical issues such as the computational complexity and storage efficiency of traditional unbiased estimation algorithms have yet to be systematically addressed. Furthermore, discussions on alternative biased estimation methods mainly focus on theoretical comparisons, lacking systematic experimental designs and empirical results across different application scenarios.

Future research needs to delve further, better and wider. First, a more comprehensive theory that can synthetically combine the merit of traditional unbiased estimation method theory and the advantage of modern machine learning algorithms is needed. In addition, priority needs to be put on designing efficient parallel computing algorithm which can help fundamentally deal with scalability problems and computation efficiency issue of the traditional approach in the age of big data. What's more, carry out more application-oriented empirical research and test via practical data experiment whether or not different types of estimations have similar power when making statements about real-world phenomena.

This study lays an important foundation for the theoretical development and methodological innovation of unbiased estimation, yet the field still presents vast exploration potential. Subsequent research should focus on overcoming the limitations of existing methodologies and, through the dual drivers of theoretical innovation and methodological improvement, continuously advance the innovative development and practical application of statistical estimation theory in the new era. Particularly in emerging fields such as artificial intelligence and big data analytics, the interdisciplinary integration of unbiased estimation theory with other disciplines will give rise to numerous research topics of significant theoretical and practical importance.

References

- [1] Shao, J.: *Mathematical Statistics*. 2nd edn. Springer, Berlin (2003). doi: 10.1007/b97553
- [2] WG.: The distribution of quadratic forms in a normal system, with applications to the analysis of covariance. *Math. Proc. Cambridge Philos. Soc.* 30(2), 178–191 (1934). doi: 10.1017/S0305004100016595
- [3] Lehmann, E.L., Casella, G.: *Theory of Point Estimation*. 2nd edn. Springer, New York (1998). doi: 10.1007/B98854
- [4] Durrett, R.: *Probability: Theory and Examples*. 4th edn. Cambridge University Press, Cambridge (2010)
- [5] Cramér, H.: *Mathematical Methods of Statistics*. Princeton University Press, Princeton (2016)
- [6] Casella, G., Berger, R.: *Statistical Inference*. 2nd edn. Chapman and Hall/CRC, Boca Raton (2024). doi: 10.1201/9781003456285
- [7] Montgomery, D.C.: *Introduction to Statistical Quality Control*. 8th edn. John Wiley & Sons, Hoboken (2020)
- [8] Lovric, M. (ed.): *International Encyclopedia of Statistical Science*. Springer, Berlin (2011). doi: 10.1007/978-3-642-04898-2
- [9] Ferguson, T.S.: *A Course in Large Sample Theory*. 2nd edn. Chapman and Hall/CRC, Boca Raton (2017). doi: 10.1201/9781315136288
- [10] Venables, W.N., Morris, K.W.: A connection between minimum variance unbiased estimation and maximum likelihood. *Am. Stat.* 43(2), 118–121 (2012)
- [11] Qiu, M., Liu, B.: Multi-sensor adaptive weighted data fusion based on biased estimation. *Sensors* 24(11), 1–20 (2024)

(Open Access Text)

Springer Nature will add OpenAccess standard text here during typesetting. Springer Nature will add OpenAccess standard text here during typesetting. Springer Nature will add OpenAccess standard text here during typesetting. Springer Nature will add OpenAccess standard text here during typesetting. Springer Nature will add OpenAccess standard text here during typesetting. Springer Nature will add OpenAccess standard text here during typesetting.

Springer Nature will add OpenAccess standard text here during typesetting. Springer Nature will add OpenAccess standard text here during typesetting. Springer Nature will add OpenAccess standard text here during typesetting. Springer Nature will add OpenAccess standard text here during typesetting. Springer Nature will add OpenAccess standard text here during typesetting. Springer Nature will add OpenAccess standard text here during typesetting.

Springer Nature will add OpenAccess standard text here during typesetting.