Robust estimations from distribution structures: V. Non-asymptotic

Tuobang Li[a,b,c,1]

[a]Technion-Israel Institute of Technology, Haifa 32000, Israel; [b]Guangdong Technion-Israel Institute of Technology, Shantou 515063, China; [c]University of California, Berkeley, CA 94720

This manuscript was compiled on February 17, 2024

Due to the complexity of order statistics, the finite sample bias of robust statistics is generally not analytically solvable. While the Monte Carlo method can provide approximate solutions, its convergence rate is typically very slow, making the computational cost to achieve the desired accuracy unaffordable for ordinary users. In this paper, we propose an approach analogous to the Fourier transformation to decompose the finite sample structure of the uniform distribution. By obtaining a set of sequences that are simultaneously consistent with a parametric distribution for the first four sample moments, we can approximate the finite sample behavior of robust estimators with significantly reduced computational costs. This article reveals the underlying structure of randomness and presents a novel approach to integrate two or more assumptions.

finite sample bias | order statistics | variance reduction | Monte Carlo study | uniform distribution

In the early nineteenth century, Bessel deduced the unbiased sample variance and found it has a correction term of $\frac{n}{n-1}$. Later, Cramér (1) in his classic textbook *Mathematical Methods of Statistics* deduced unbiased sample central moments with a linear time complexity. However, apart from the mean and central moments, the finite sample behavior of nearly all other estimators depends on the underlying distribution and lacks a simple non-parametric correction term. For example, the simplest robust estimator, the median, exhibits a highly complex finite sample behavior. If n is odd, $E\left[median_n\right] = \int_{-\infty}^{\infty} \left(\frac{n+1}{2}\right)\binom{n}{\frac{n}{2}-\frac{1}{2}} F\left(x\right)^{\frac{n}{2}-\frac{1}{2}}\left[1-F(x)\right]^{\frac{n}{2}-\frac{1}{2}} f(x)xdx$ (2), where $F(x)$ and $f(x)$ represent the cumulative distribution function (cdf) and probability density function (pdf) of the assumed distribution, respectively. For the exponential distribution, the above equation is analytically solvable, yielding $E\left[median_n\right] = \frac{2^{-n-1}(n+1)\binom{n}{\frac{n-1}{2}}\left(H_n - H_{\frac{n-1}{2}}\right)\Gamma\left(\frac{n+1}{2}\right)\sqrt{\pi}}{\lambda\Gamma\left(\frac{n}{2}+1\right)}$, where H_n denotes the nth Harmonic number, Γ represents the gamma function, and λ stands for the scale parameter of the exponential distribution. However, for distributions with more complex pdfs, such equations are generally unsolvable. Another widely used exact finite sample bias correction is the factor for unbiased standard deviation in the Gaussian distribution, which can be deduced using Cochran's theorem (3). For more complex estimators, writing their exact finite-sample distribution formulas becomes challenging. In 2013, Nagatsuka, Kawakami, Kamakura, and Yamamoto derived the exact finite-sample distribution of the median absolute deviation, which consists of four cases, each with a lengthy formula (4). In such cases, even obtaining a numerical solution is challenging (2, 4). So, Monte Carlo simulation is currently the only practical choice for estimating finite sample corrections. However, the computational cost of Monte Carlo simulation is often too high to be processed on a typical PC. For example, for median absolute deviation, Croux and Rousseeuw (1992) provided correction factors with a precision of three decimal places for $n \leq 9$ using 200,000 pseudorandom Gaussian sample (5). Hayes (2014) reported correction factors for $n \leq 100$ using 1 million pseudorandom samples for each value of n to ensure the accuracy to four decimal places (6). Recently, Akinshin (2022) (7) presented correction factors for $n \leq 3000$ using 0.2-1 billion pseudorandom Gaussian samples. His result suggest that, for the median absolute deviation, finite sample bias correction is required to ensure a precision of three decimal places when the sample size is smaller than 2000. This highlights the importance of finite sample bias correction. However, since different correction factors are required for different parametric assumptions, the computational cost of addressing all possible cases in the real world becomes significant, especially for complex models.

In addition to computational challenges, there exists an inherent difficulty in dealing with randomness. The theory of probability provides a framework for modeling and understanding random phenomena. However, the practical implementation of these models can be challenging, as discussed, and their complexity greatly hinders our comprehension. The quality of randomness can significantly impact the validity of simulation results, and a deeper understanding of randomness may offer a more effective and cost-efficient solution. The purpose of this brief report is to demonstrate that the finite sample structure of uniform random variables can be decomposed using a few well-designed sequences with high accuracy. Furthermore, we show that the computational cost of estimating finite sample bias from a Monte Carlo study can be

Significance Statement

Most contemporary statistics theories focus on asymptotic analysis due to its tractability and simplicity. Non-asymptotic statistics are crucial when dealing with small or moderate sample sizes, which is often the case in practice. In situations where analytical results are difficult or impossible to obtain, Monte Carlo studies serve as a powerful tool for addressing non-asymptotic behavior. However, these studies can be computationally expensive, particularly when high precision is required or when the statistical model demands significant computational time. Here, we propose calibrated Monte Carlo study that aims to approximate the randomness structures using a small set of sequences. This approach sheds light on understanding the general structure of randomness.

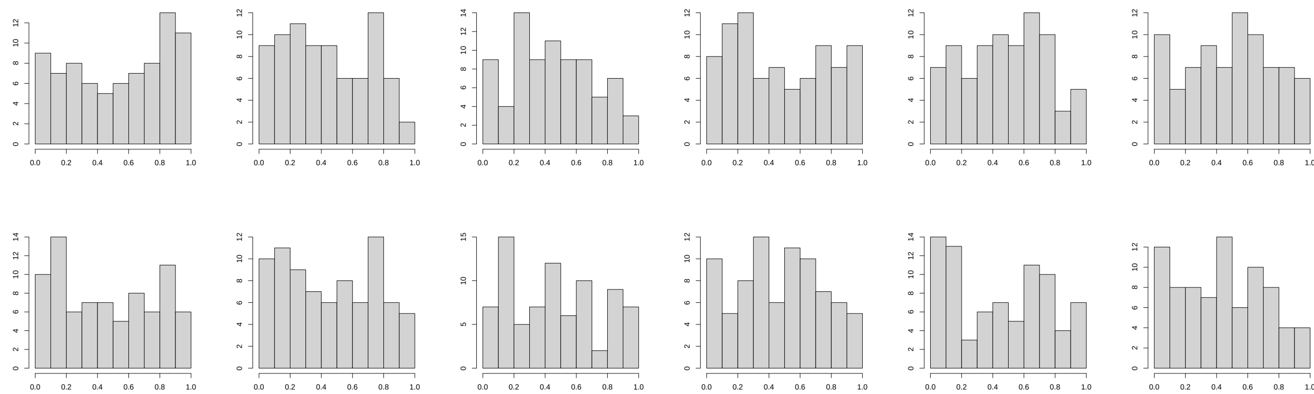

Fig. 1. The frequency histograms of pseudo-random sequences on the interval [0,1] with size 80.

dramatically improved by obtaining a set of sequences that are simultaneously consistent with a parametric distribution for sample central moments.

Decomposing the finite sample structure of uniform distribution

Any continuous distribution can be linked to the uniform distribution on the interval [0, 1] through its quantile function. This fundamental concept in Monte Carlo study implies that understanding the finite sample structure of uniform random variables can be leveraged to understand the finite sample structure of any other continuous random variable through the quantile transform. The Glivenko–Cantelli theorem (8, 9) ensures the almost-sure convergence of the empirical distribution function to the true distribution function. However, the individual empirical distribution often deviates significantly from the asymptotic distribution even when the sample size is not small (Figure 1, sample size is 80), which cause finite sample biases of common estimators. Let μ, μ_2, ..., $\mu_{\mathbf{k}}$ denote the first $\mathbf{k}$ central moments of a probability distribution. According to the unbiased sample central moment (1), the expected value of the sample central moment, $m_{\mathbf{k}} = \frac{1}{n}\sum_{\mathbf{k}=1}^{n}(x_{\mathbf{k}} - \bar{x})^{\mathbf{k}}$, can be deduced, denoted as $E[m_{\mathbf{k}}]$. Let $S = \{sequence[i] | i \in \mathbb{N}\}$ be a set of number sequences ranging from 0 to 1, where $sequence[i]$ represents the ith sequence in the set, and $\mathbb{N}$ is the set of natural numbers, with $i \leq N$. Transform every number in S using the quantile function of a parametric distribution, PD. The transformed sequences can be denoted as S_{PD}. Denote the set of the $\mathbf{k}$th sample central moments for these transformed sequences as $M_{\mathbf{k}} = \{m_{\mathbf{k},i} | i \in \mathbb{N}\}$. S is consistent with PD for all $m_{\mathbf{k}}$ when $\mathbf{k} \leq k$, if and only if the following system of linear equations is consistent,

$$\begin{cases} m_{1,1}w_1 + \ldots + m_{1,i}w_i + \ldots + m_{1,N}w_N = E[m_1] \\ \ldots \\ m_{\mathbf{k},1}w_1 + \ldots + m_{\mathbf{k},i}w_i + \ldots + m_{\mathbf{k},N}w_N = E[m_{\mathbf{k}}] \\ \ldots \\ m_{k,1}w_1 + \ldots + m_{k,i}w_i + \ldots + m_{k,N}w_N = E[m_k] \\ w_1 + \ldots + w_i + \ldots + w_N = 1 \end{cases}, \text{ where}$$

w_1, ..., w_i, ..., w_N are the unknowns of the system, with $N \geq k+1$. w_1, ..., w_i, ..., w_N can be determined using a typical constraint optimization algorithm. The Monte Carlo study can be seen as a special case when $w_1 = \ldots = w_i = \ldots = w_N$, and the sequences in S are all random number sequences. The strong law of large numbers (proven by Kolmogorov in 1933) (10) ensures that in this case, when the number of sequences $N \to \infty$ or when the sample size $n \to \infty$, the above system of linear equations is always consistent.

Low-discrepancy sequences are commonly used as a replacement of uniformly distributed random numbers to reduce computational cost. When considering a sequence to approximate the structure of uniform random variables, the most natural choice is the arithmetic sequence, denoted as $\{x_i\}_{i=1}^{n} = \left\{\frac{i}{n+1}\right\}_{i=1}^{n}$. However, the arithmetic central moments estimated from the arithmetic sequence for the Gaussian distribution differ significantly from their expected values (Figure 2A). The arithmetic sequence lacks the variability of true random samples which produce additional biases for even order moments. The beta distribution is defined on the interval (0, 1) in terms of two shape parameters, denoted by α and β. When $\alpha = \beta$, the beta distribution is symmetric. To better replicate the features of uniform random variables, we introduced beta distributions with a variety of parameters. The arithmetic sequences were transformed by the quantile functions of these beta distributions to form beta-sequences, resulting in sequences that are U-shape ($\alpha = \beta = 0.547$), left-skewed ($\alpha = 46.761$, $\beta = 20.108$), right-skewed ($\alpha = 20.108$, $\beta = 46.761$), monotonic decreasing ($\alpha = 0.478$, $\beta = 38.53$), monotonic increasing ($\alpha = 38.53$, $\beta = 0.478$), their left-skewed self-mixtures ($\alpha = \beta = 0.369$, $\alpha = \beta = 18.933$), their right-skewed self-mixtures ($\alpha = \beta = 0.369$, $\alpha = \beta = 18.933$), their left-skewed mixture with the arithmetic sequence ($\alpha = \beta = 0.328$), their right-skewed mixture with the arithmetic sequence ($\alpha = \beta = 0.328$) (Figure 2B). Besides beta sequences with a U-shape, other sequences are paired so an additional constraint is set to ensure equal weight for each pair. Besides these 9 sequences and arithmetic sequences, a pseudo-random sequence is introduced to further approximate the structure and avoid inconsistent scenarios. Finally, a complement sequence is introduced which if combining all the sequences with corresponding weights, the overall sequence is nearly uniform.

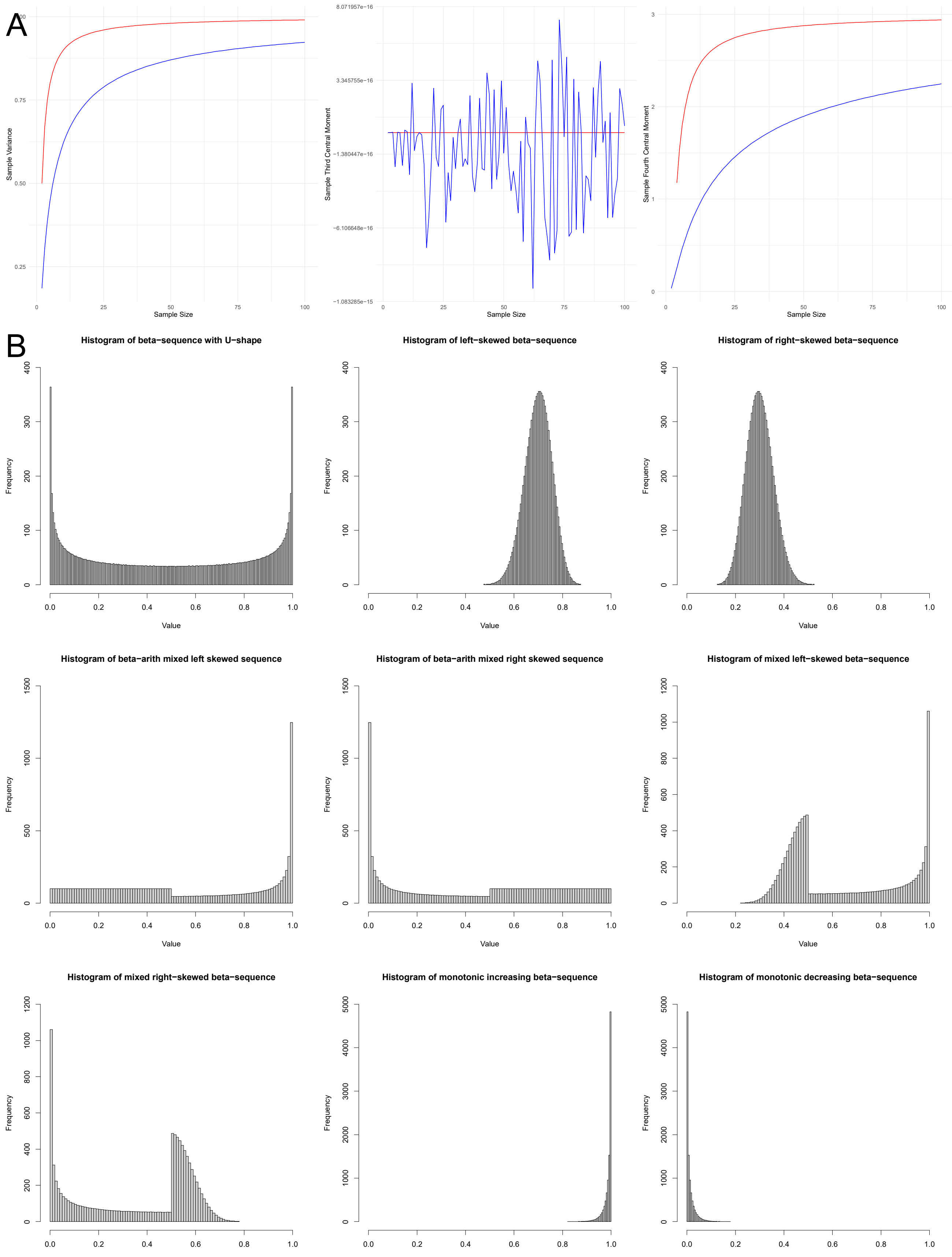

Fig. 2. A. The first four sample central moments for the Gaussian distribution are plotted over a sample size ranging from 2 to 100. The red lines represent the expected values, while the blue lines depict the values estimated from the arithmetic sequences. B. The histograms of different beta sequences, their self-mixtures, and mixtures with arithmetic sequences.

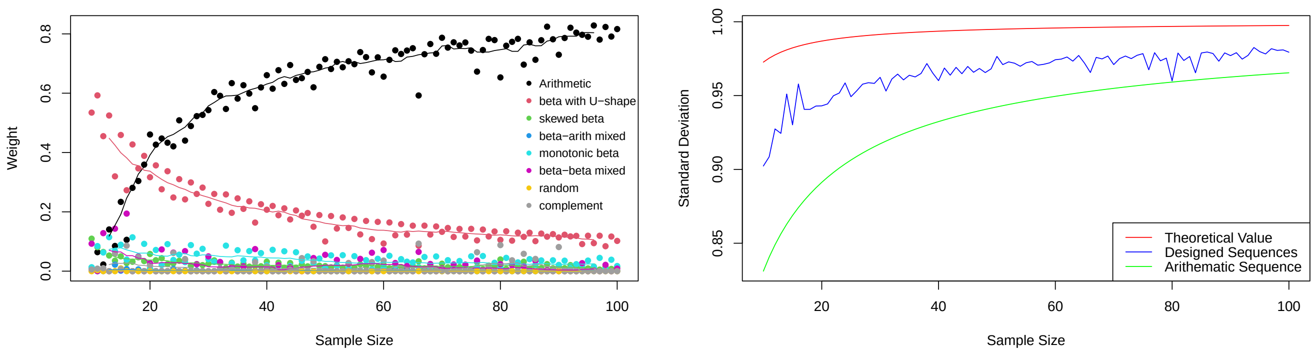

Fig. 3. The first plot shows the weights assigned to different sequences as the sample size increases. The second plot depicts the sample standard deviations estimated from designed and arithmetic sequences and compares them to the true values. The designed sequences were repeated 10 times to reduce the variation due to the random sequences.

Results

The most surprising result in this article is that, by carefully selecting/designing sequences in S, even when N and n are very small, e.g., less than 20, the above system of linear equations can still be consistent, while the weight assigns to the random and complement sequences are extremely small (<0.01 on average). Using just 12 sequences, when $n = 10$, the constraint optimization algorithm can assign weights to all these sequences with errors less than 10^{-10}. This means that technically, these sequences are consistent with the Gaussian distribution for the first four moments. More importantly, the findings suggest that when the sample size is small, the beta sequence with a U-shape accounts for approximately 50-60% of the finite sample properties of uniform random variables, while arithmetic, monotonic beta, beta-beta mixed, skewed beta distributions each contribute about 2-10% (Figure 3). As the sample size grows, as expected, the weight of the arithmetic sequence increases and dominants while the weights of other sequences gradually decrease. However, the beta sequence with a U-shape still still holds about 10% weight even when the sample size is 100 (Figure 3).

The obtained weights can be used to estimate the finite sample behaviour of other related estimators, such as the standard deviation and median absolute deviation for the Gaussian distribution. We found that by using the 12 well-designed sequences, the performance is much better than the arithmetic sequence (Figure 3). To further increase precision, we adopted a stochastic method. We pseudo-randomly generated twelve sequences and evaluated their efficacy in approximating the finite sample structure of uniform random variables by solving the above system of linear equations for the first four moments. Sequences that met the predetermined accuracy threshold (error less than 10^{-5}) were retained, while those that did not meet the requirement were discarded in favor of a new set. Upon identifying twenty qualified sets, these sets were applied to assess the finite sample biases in other estimators for the Gaussian distribution. The outcomes indicate that using merely fifty sets of sequences, totaling 600 sequences, which can be executed on a standard PC in a negligible amount of time, achieves a precision of approximately 0.005 for the standard deviation and median absolute deviation. In contrast, attaining the same level of precision using classic Monte Carlo methods would require roughly 0.1 million pseudo-random samples.

Data and Software Availability

All data are included in the brief report and SI Dataset S1. All codes have been deposited in GitHub.

1. H Cramér, *Mathematical methods of statistics.* (Princeton university press) Vol. 43, (1999).
2. HA David, HN Nagaraja, Order statistics, third edition in *Wiley Series in Probability and Statistics.* (2003).
3. WH Holtzman, The unbiased estimate of the population variance and standard deviation. *The Am. J. Psychol.* **63**, 615–617 (1950).
4. H Nagatsuka, H Kawakami, T Kamakura, H Yamamoto, The exact finite-sample distribution of the median absolute deviation about the median of continuous random variables. *Stat. & Probab. Lett.* **83**, 999–1005 (2013).
5. C Croux, PJ Rousseeuw, Time-efficient algorithms for two highly robust estimators of scale. (1992).
6. K Hayes, Finite-sample bias-correction factors for the median absolute deviation. *Commun. Stat. - Simul. Comput.* **43**, 2205–2212 (2014).
7. A Akinshin, Finite-sample bias-correction factors for the median absolute deviation based on the harrell-davis quantile estimator and its trimmed modification (2022).
8. V Glivenko, Sulla determinazione empirica delle leggi di probabilita. *Gion. Ist. Ital. Attauri.* **4**, 92–99 (1933).
9. FP Cantelli, Sulla determinazione empirica delle leggi di probabilita. *Giorn. Ist. Ital. Attuari* **4** (1933).
10. A Kolmogorov, Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.* **4**, 83–91 (1933).