Proceedings of CONF-APMM 2026 Symposium: Applied Mathematics and Quantitative Modeling
DOI: 10.54254/2753-8818/2026.PU38792

Application of Probability Theory and Mathematical Statistics in Big Data Analysis

Zihan Tang

Sentinel Secondary School, West Vancouver, Canada
Alextang0913@gmail.com

Abstract. With the rapid growth of the Internet, the Internet of Things, and digital platforms, large-scale data analysis has become increasingly important in many practical fields. However, large datasets still contain randomness, variability, and uncertainty, which require appropriate statistical methods for reliable interpretation. This study examines the application of probability theory and mathematical statistics in big data analysis from the perspectives of sampling distributions, numerical estimation, and parameter inference. Three statistical simulation experiments were carried out using Python to examine the Central Limit Theorem, Monte-Carlo estimation, and confidence-interval analysis over a range of sample-size conditions. The results show that increasing the sample size significantly improves statistical stability. The standard deviation of sample means decreased from 0.6337 to 0.0906, the mean absolute Monte Carlo estimation error declined from 0.12510 to 0.00394, and the average width of the 95% confidence interval decreased from 7.3884 to 0.5566. These findings indicate that probability theory and mathematical statistics remain fundamental tools for describing uncertainty, evaluating estimation accuracy, and supporting reliable analysis in large-scale data environments.

Keywords: Probability Theory, Mathematical Statistics, Big Data Analysis, Monte Carlo Method, Statistical Inference

1. Introduction

With the continuous development of the Internet, the Internet of Things (IoT), and various digital platforms, the volume of data generated in industrial production and daily activities has expanded rapidly. Data sources have evolved from traditional structured information to encompass diverse forms such as transaction records, device operational data, online behavior, and textual content [1]. While this growth in data volume provides a richer foundation for analysis, it also introduces challenges such as significant data volatility, complex variable relationships, and inherent result uncertainty. Basic data curation and heuristic comparisons are often inadequate for fully revealing latent underlying patterns in data. Consequently, extracting representative features from vast amounts of information and reasonably interpreting stochastic variations have become critical issues in big data analysis [2].

Probability theory and mathematical statistics offer a robust mathematical framework for addressing these issues. Probability theory focuses on the study of random phenomena and their

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.PU38792

governing patterns, whereas mathematical statistics emphasizes analyzing and inferring population characteristics from available information [3]. Fundamental techniques—such as the mean, variance, probability distributions, sampling, and interval estimation—enable description of data from various perspectives, including central tendency, range of fluctuation, and potential trends. These methods remain effective even as data volumes scale up, indeed, with larger datasets, probabilistic and statistical methods are essential for assessing result stability, identifying significant differences between samples, and determining the reliability of analytical conclusions [4].

Existing research demonstrates that probabilistic and statistical methods are widely applied in fields such as data mining, economic analysis, business decision-making, and risk analysis [5, 6]. Studies suggest that large-scale data analysis requires not only computational tools to enhance processing efficiency but also statistical methods to interpret inter-data relationships and result uncertainties. In practical applications, sampling analysis alleviates the computational burden of processing entire datasets, statistical estimation enables inferences about the population based on sample characteristics, and the Monte Carlo method facilitates the study of complex patterns through extensive random simulations [7]. Furthermore, research has utilized probability distributions and statistical inference to analyze economic and business data, providing valuable insights for trend assessment and risk control [8]. These studies demonstrate that probability theory and mathematical statistics remain vital theoretical tools for big-data analysis, however, the statistical behaviours exhibited by different methods under varying sample sizes warrant further elucidation grounded in empirical results.

Against this background, this study aims to clarify how fundamental principles of probability theory and mathematical statistics can be used to characterize uncertainty and evaluate estimation stability in data-intensive analytical settings. Specifically, the study will address three questions: how sample size affects the sampling distribution of the mean, how the number of random samples influences Monte Carlo estimation accuracy, and how sample size affects the precision of parameter estimation and confidence intervals. To address these questions, the study first reviews the relevant theoretical foundations and then conducts three controlled Python-based simulations as illustrative experiments. These simulations are intended to demonstrate fundamental statistical mechanisms under transparent conditions rather than to provide empirical validation using a specific real-world big-data dataset. By linking statistical theory with controlled numerical experiments, this study seeks to provide a clearer understanding of the role of probability and mathematical statistics in uncertainty characterization, estimation assessment, and reliable reasoning in large-scale data analysis.

2. Theoretical foundations of probability theory and mathematical statistics

2.1. Random variables and probability distributions

In practical data analysis, observations pertaining to the same phenomenon may vary across different times and distinct subjects. Such variation often contains a certain degree of randomness. To describe uncertain phenomena quantitatively, probability theory usually uses a random variable to represent the possible numerical outcomes of a random experiment [9, 10]. According to the characteristics of their values, random variables can generally be divided into discrete random variables and continuous random variables. The possible values of a discrete random variable can be listed individually, while a continuous random variable may take any value within a certain interval. Probability distributions can further describe the likelihood of different outcomes and therefore provide a basis for subsequent data analysis.

DOI: 10.54254/2753-8818/2026.PU38792

For a continuous random variable, its variation pattern can usually be described by a Probability Density Function (PDF). A probability distribution can not only indicate the possible range of the data but also reflect how the observations are concentrated in different regions. In practical analysis, expectation and variance are two commonly used statistical measures. For a continuous random variable X with probability density function f(x), its expectation can be expressed as

$$E\left(X\right)=\int_{-\infty}^{+\infty}xf\left(x\right)dx \tag{1}$$

where E(X) represents the average level shown by the random variable over repeated observations. Variance is used to measure the degree of fluctuation of a random variable around its mean and can be written as

$$Var(X)=E[(X-\mu)^{2}] \tag{2}$$

where μ denotes the expectation of the random variable. A smaller variance generally indicates that the data are more concentrated, while a larger variance suggests greater dispersion. Therefore, in large-scale data analysis, expectation can help describe the overall level of the data, while variance reflects the degree of variation. Together, they form an important basis for describing the main characteristics of data.

2.2. Law of large numbers and central limit theorem

As the amount of data increases, random fluctuations in individual observations do not disappear completely, but the overall pattern formed by a large number of samples often becomes more stable. The Law of Large Numbers (LLN) states that when the sample size is sufficiently large, the sample mean converges toward the true population mean[11, 12]. This principle provides an important theoretical basis for using sample information to understand population characteristics and also shows that increasing the sample size can reduce the influence of accidental variation on the final result.

Based on this principle, the Central Limit Theorem (CLT) further explains the distribution of sample means. Under certain conditions, even when the original population does not follow a normal distribution, the sampling distribution of the sample mean converges toward a normal distribution as the sample size increases. Its standardized form can be expressed as

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\rightarrow N(0,1) \tag{3}$$

where $\bar{X}$ is the sample mean, μ and σ represent the population mean and population standard deviation, respectively, and n is the sample size. This expression indicates that as n increases, the fluctuation of the sample mean generally becomes smaller. The Central Limit Theorem is also an important theoretical foundation for many statistical methods applied to large-sample problems and provides support for mean estimation, error analysis, and interval estimation.

2.3. Parameter estimation and confidence intervals

In practical analysis, the true characteristics of a population are usually not directly available. Therefore, sample information is often used to estimate unknown population parameters. Parameter

DOI: 10.54254/2753-8818/2026.PU38792

estimation refers to the process of using statistical characteristics obtained from a sample to estimate unknown population parameters. For example, the sample mean may be used to estimate the population mean, whereas the sample standard deviation quantifies the degree of variation within the population.

Although a single estimated value can provide an approximate description of the population, it cannot fully indicate the uncertainty associated with the estimation result. For this reason, mathematical statistics commonly uses a Confidence Interval (CI) to provide a possible range for an unknown parameter. When the population standard deviation is unknown, the confidence interval for the population mean can be expressed as

$$\bar{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \tag{4}$$

where s represents the sample standard deviation and $t_{\alpha/2,n-1}$ is the critical value corresponding to a given confidence level. The formula shows that, when other conditions remain unchanged, an increase in the sample size n reduces the standard error and therefore narrows the confidence interval. This means that larger samples usually provide more stable parameter estimates and reduce the uncertainty involved in statistical analysis.

2.4. Monte Carlo Method

The Monte Carlo Method is a numerical approach based on random sampling. Its basic idea is to perform a large number of repeated random experiments, observe the aggregate pattern of the outcomes, and use these results to obtain an approximate value for the target quantity. Compared with directly solving complex mathematical expressions, this method can be useful when a problem is difficult to solve analytically. It has therefore been widely used in risk analysis, numerical calculation, and uncertainty analysis.

The estimation of π provides a simple example. A large number of random points can be generated inside a square with side length 2, and each point can then be checked to determine whether it lies inside a unit circle. If the random points are distributed sufficiently evenly, the ratio of points inside the circle to the total number of points gradually approaches the ratio of the area of the circle to the area of the square. Thus, π can be estimated by

$$\hat{\pi} = 4\frac{N_{inside}}{N} \tag{5}$$

where N_{inside} represents the number of points located inside the unit circle and N represents the total number of random points. As the number of random samples increases, the estimation result generally becomes more stable, and the error usually decreases at an approximate rate of $N^{-1/2}$. This shows that the number of random samples is an important factor affecting the accuracy of Monte Carlo estimation and provides a direct way to understand the stability of random calculations under large-scale data conditions.

3. Application of probability theory and mathematical statistics in big data analysis

3.1. Probability distribution modeling

In big data analysis, different data types often exhibit distinct variation patterns, relying solely on simple metrics — such as averages or maximum values — fails to fully capture the underlying

DOI: 10.54254/2753-8818/2026.PU38792

characteristics of the dataset as a whole. Probability distributions describe the likelihood of data falling within specific ranges from a holistic perspective, consequently, they are frequently used to analyze central tendency, dispersion, and anomalous fluctuations within large datasets. For instance, in user behavior analysis, metrics such as visit counts, dwell times, and purchase frequencies vary significantly among users, in financial data, rates of return fluctuate with changing market conditions, and in industrial production, parameters like equipment temperature, pressure, and operating time are influenced by a multitude of factors. Modeling these data using probability distributions allows for a more intuitive observation of their behavioral characteristics and establishes a foundation for subsequent analysis.

It is important to note that no single, universally applicable probability model exists for all data types. Some datasets may approximate a normal distribution, while others exhibit pronounced skewness or "long-tail" characteristics. Therefore, in practice, the appropriate model should be selected based on the data's range, distributional shape, and specific research objectives, rather than simply assuming a uniform distribution pattern across all datasets. Selecting an appropriate probability distribution enables more accurate characterisation of data features and minimises analytical bias arising from misspecified model selection.

3.2. Sampling and statistical inference

Although big data environments involve massive volumes of data, statistical inference remains highly relevant. Processing an entire large-scale dataset for every analysis can increase computational time and storage demands, potentially reducing overall efficiency. Consequently, provided the sample is sufficiently representative, obtaining a subset of data through sampling—and using that sample to infer population characteristics—remains an effective approach. Judicious sampling reduces the computational burden while preserving key information, thereby enhancing the efficiency of the analysis process.

On the other hand, a large data volume does not by itself guarantee representativeness or eliminate uncertainty. Real-world big-data sources are often collected through convenience mechanisms rather than simple random sampling. For example, platform data may disproportionately represent active users while excluding non-users or less-active groups. It is therefore important to distinguish sampling error from systematic selection bias. Sampling error refers to random fluctuation caused by observing a sample rather than the entire population and can generally be reduced by increasing the sample size under an appropriate sampling design. In contrast, systematic selection bias arises from the data-collection mechanism itself and does not automatically disappear as the sample size increases. Missing values, measurement errors, duplicate records, and noise may introduce additional uncertainty. Consequently, the reliability of big-data analysis should be evaluated not only by the amount of data but also by the representativeness and quality of the data source. Statistical measures can quantify stochastic variability, but they cannot by themselves correct systematic selection bias.

3.3. Uncertainty quantification

In data analysis, providing only a single final value often fails to fully convey the reliability of the results. For instance, two sets of results might share the same mean yet exhibit vastly different ranges of data fluctuation. Relying solely on the mean risks overlooking the underlying uncertainty. Uncertainty Quantification (UQ) addresses this by focusing on the potential range of fluctuation in

DOI: 10.54254/2753-8818/2026.PU38792

analysis results and characterizing this uncertainty using metrics such as variance, standard deviation, margin of error, and confidence intervals.

The confidence interval is a common method in this context. Unlike a single point estimate, a confidence interval provides a range of values likely to contain the population parameter, enabling analysts to more intuitively assess the stability of the estimate. Generally, assuming other conditions remain constant, increasing the sample size reduces the impact of individual random factors on the estimate, typically narrowing the range of fluctuation. Consequently, in large-scale data analysis, reporting both the estimated value and the associated uncertainty range offers a better understanding of the results than a single figure alone, providing richer information for subsequent decision-making.

3.4. Application of monte carlo methods in large-scale analysis

The Monte Carlo method derives approximate solutions to problems through extensive random sampling and repeated computation, it is particularly well-suited for addressing problems influenced by multiple random factors. In financial risk analysis, large-scale stochastic scenarios can be used to observe the range of potential losses under varying market conditions, in engineering reliability research, randomly varying equipment parameters allow for the analysis of system failure probabilities under different conditions, and in numerical computation, stochastic experiments can be employed to progressively approximate target values when precise analytical solutions are difficult to obtain directly.

This method integrates well with large-scale data analysis, since the stability of its results is closely tied to the size of the random sample. With fewer samples, individual stochastic outcomes can significantly influence the final estimate, however, as the sample size increases, such random fluctuations typically diminish, leading to more stable computational results. Consequently, the Monte Carlo method not only addresses complex stochastic problems but also provides an intuitive illustration of the relationship between sample size and estimation error. Building on this characteristic, the following section analyzes how estimation error and result stability evolve as the sample size increases, drawing upon computational results obtained at various scales.

4. Simulation experiments and results

4.1. Experimental design

To examine how variations in sample size affect statistical outcomes, three independent experiments were performed to investigate the sampling distribution of sample means, random estimation error, and the stability of interval estimation. All calculations were performed in Python using NumPy, Pandas, SciPy, and Matplotlib for numerical processing and visualization. A fixed random seed of 42 was used to ensure reproducibility. The main settings of the three experiments are summarized in Table 1.

Table 1. Main settings of the three statistical experiments

Experiment	Population / Target	Sample Size	Repetitions	Main Evaluation
Central Limit Theorem	Exponential distribution, population size = 100000	10, 30, 100, 500	5000	Distribution and standard deviation of sample means

DOI: 10.54254/2753-8818/2026.PU38792

Table 1. (continued)

Monte Carlo Method	Estimation of pi	100-100000	30	Mean absolute estimation error
Confidence Interval	Normal population, mu = 50, sigma = 10	30, 100, 500, 1000, 5000	500	Mean estimate and 95% confidence interval

The three experiments approach the problem from different perspectives: distributional behavior, random numerical error, and parameter-estimation accuracy. The purpose is not to increase model complexity, but to determine whether changes in sample size produce stable and interpretable statistical effects under controlled conditions.

It should be noted that these experiments are deliberately simplified, controlled simulations. They are designed to isolate and illustrate the statistical effects of sample size rather than replicate the full complexity of real-world big data applications.

4.2. Central limit theorem simulation

The first experiment adopted an exponential distribution with pronounced positive skewness as the population and repeatedly computed sample means across different sample sizes. This choice avoids the prior advantage of starting from a normal population and makes changes in the sampling distribution easier to observe. Figure 1 presents the results for n = 10, 30, 100, and 500.

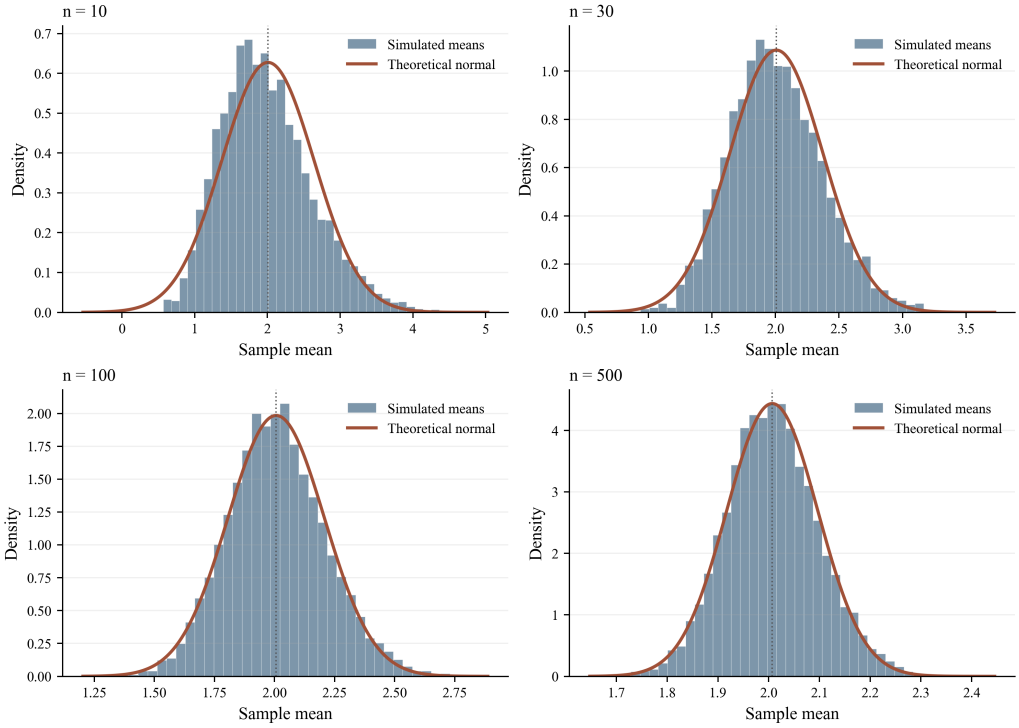


Figure 1. Distribution of sample means under different sample sizes

When n = 10, the distribution of sample means still shows noticeable skewness and a relatively wide spread. As the sample size increases, the distribution becomes increasingly concentrated

DOI: 10.54254/2753-8818/2026.PU38792

around its center, and clear symmetry is observed at n = 100 and n = 500. Quantitatively, the standard deviation of the sample means decreases from 0.6337 at n = 10 to 0.0906 at n = 500, a reduction of approximately 85.7%. Over the same range, the theoretical standard error decreases from about 0.6358 to 0.0899, closely matching the empirical results.

These results indicate that larger samples substantially reduce the influence of random variation on the sample mean. More importantly, although the original population is non-normal, the distribution of the sample mean becomes progressively closer to a normal shape as the sample size grows, which is consistent with the Central Limit Theorem.

4.3. Monte carlo convergence analysis

The second experiment estimated pi from the proportion of random points falling inside a unit circle and compared the estimation error across different numbers of random samples. Because a single simulation may be affected by considerable random fluctuation, each sample size was repeated 30 times. The mean absolute error across repeated runs was therefore adopted as the primary performance metric.

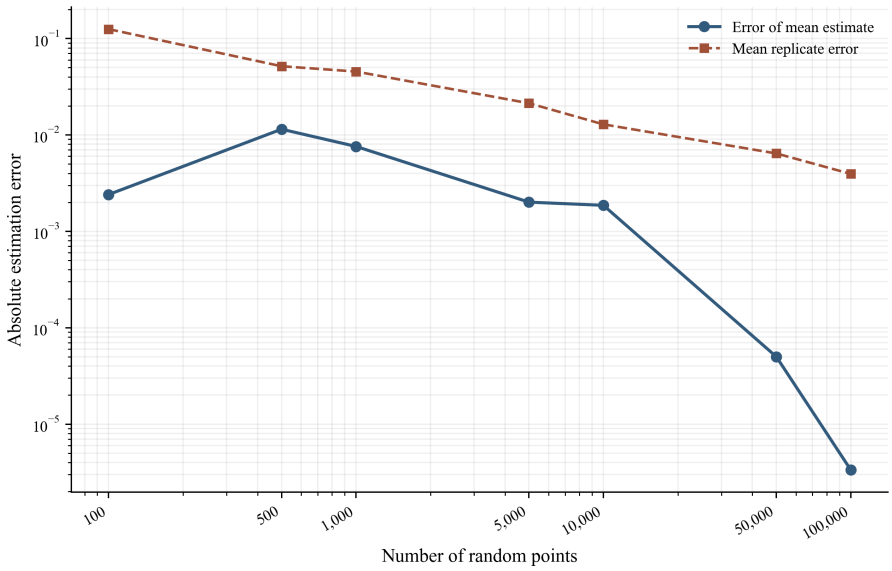


Figure 2. Monte Carlo estimation error versus sample size

Table 2. Mean absolute error of Monte Carlo estimation

Sample Size	Mean Absolute Error
100	0.12510
500	0.05160
1000	0.04540
5000	0.02130
10000	0.01290
50000	0.00642
100000	0.00394

DOI: 10.54254/2753-8818/2026.PU38792

As shown in Figure 2 and Table 2, as the number of random points increases from 100 to 100000, the mean absolute error decreases from 0.12510 to 0.00394, corresponding to an overall reduction of approximately 96.9%. More than half of the error is removed when the sample size rises from 100 to 500. Beyond 10000 samples, the error continues to fall, although the marginal reduction becomes smaller. The overall decline is compatible with the expected N^(-1/2) convergence tendency of Monte Carlo estimation.

Because Monte Carlo calculations are stochastic, strict monotonicity between every two adjacent sample sizes is not required. The mean absolute error across repeated runs is more informative than a single-run error because it better reflects the overall stability of the estimator.

4.4. Statistical estimation and confidence interval analysis

The third experiment examined the effect of sample size on mean estimation and interval width. The population mean was set to 50, and repeated samples were drawn across several sample sizes to compute the estimated mean and corresponding 95% confidence intervals.

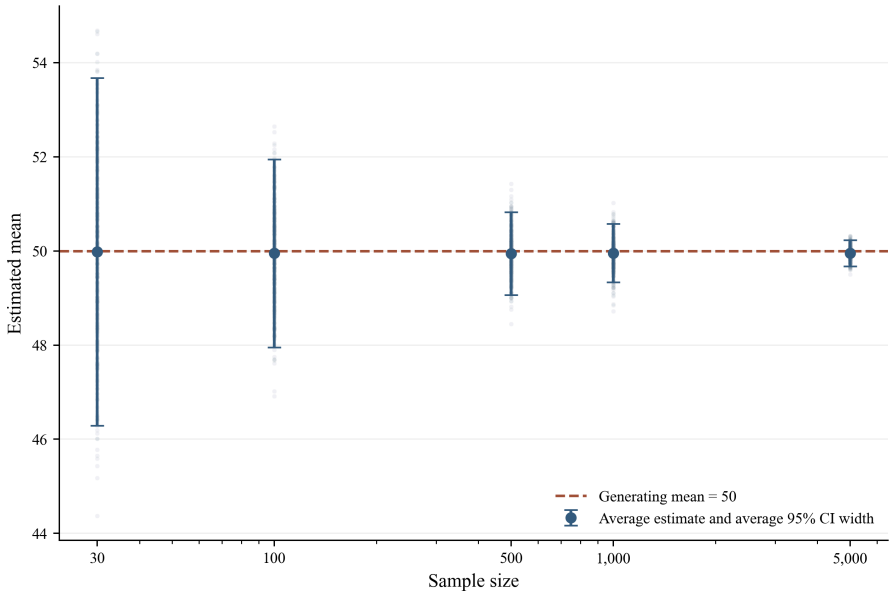


Figure 3. Mean estimation and 95% confidence intervals under different sample sizes

Figure 3 shows that the estimated means remain close to the true population mean of 50 across all sample sizes, while the intervals are visibly wider for small samples. As the sample size increases, the error bars shorten consistently, indicating a clear reduction in estimation uncertainty. The average confidence-interval width falls from 7.3884 at n = 30 to 0.5566 at n = 5000, representing a reduction of approximately 92.5%.

Across repeated experiments, the empirical coverage rate remains between approximately 93.4% and 96.4%, which is close to the nominal 95% confidence level. The small departures from exactly 95% are consistent with simulation-to-simulation random fluctuation under a finite number of repetitions (500) and should not be interpreted as evidence of systematic undercoverage or overcoverage. This result shows that increasing the sample size mainly improves estimation precision and narrows the interval rather than simply changing the confidence level itself. At n = 5000, the interval width is only about 7.5% of that observed at n = 30, and the estimate remains tightly concentrated around the population mean.

DOI: 10.54254/2753-8818/2026.PU38792

5. Conclusion

This paper investigates the application of probability theory and mathematical statistics in big data analysis. It outlines relevant theories—covering probability distributions, sampling and statistical inference, uncertainty analysis, and stochastic computation—and analyzes how statistical patterns shift across varying sample sizes. This study focuses on how variations in sample size affect the sampling distribution of the sample mean, stochastic estimation error, and the stability of parameter estimation, thereby demonstrating the practical utility of classical probabilistic and statistical methods for large- scale data processing.

The results from the three sets of experiments reveal consistent patterns. The Central Limit Theorem experiment demonstrates that as the sample size increases from 10 to 500, the sampling distribution of the sample mean becomes increasingly concentrated — with its standard deviation falling from 0.6337 to 0.0906 — and progressively exhibits characteristics approximating a normal distribution. In the Monte Carlo experiment, increasing the sample size from 100 to 100,000 reduced the mean absolute error from 0.12510 to 0.00394, indicating that expanding the scale of random sampling significantly improves the stability of estimation results. The confidence interval experiment further shows that as the sample size increases, the estimated mean consistently remains close to the population mean, while the average width of the 95% confidence interval narrows from 7.3884 to 0.5566, signifying a substantial reduction in the range of uncertainty for parameter estimation.

Overall, probability theory and mathematical statistics remain vital mathematical foundations for big data analysis. While increasing data scale can mitigate some random fluctuations, it cannot fully eliminate errors, biases, and uncertainties. This paper illustrates these patterns using fundamental statistical scenarios and does not yet address issues common to real-world large-scale datasets, such as missing data, high-dimensional features, variations in data sources, and variable correlations. Future research could incorporate actual datasets and employ methods such as Bayesian analysis, high-dimensional statistics, and statistical learning to further evaluate the applicability of probabilistic and statistical tools in complex big data environments.

References

[1] Tang, L., Li, J., Du, H., Li, L., Wu, J., & Wang, S. (2022). Big data in forecasting research: A literature review. *Big Data Research, 27*, 100289. doi:10.1016/j.bdr.2021.100289.

[2] Dasari, S., & Kaluri, R. (2023). Big data analytics, processing models, taxonomy of tools, V's, and challenges: State-of-art review and future implications. *Wireless Communications and Mobile Computing, 2023*, 3976302. doi:10.1155/2023/3976302.

[3] Gao, Y., Liu, W., Wang, H., Wang, X., Yan, Y., & Zhang, R. (2022). A review of distributed statistical inference. *Statistical Theory and Related Fields, 6*(2), 89–99. doi:10.1080/24754269.2021.1974158.

[4] Yu, J., Ai, M., & Ye, Z. (2024). A review on design inspired subsampling for big data. *Statistical Papers, 65*(2), 467–510. doi:10.1007/s00362-022-01386-w.

[5] Li, L., Lin, J., Ouyang, Y., & Luo, X. (2022). Evaluating the impact of big data analytics usage on the decision-making quality of organizations. *Technological Forecasting and Social Change, 175*, 121355. doi:10.1016/j.techfore.2021.121355.

[6] Oesterreich, T. D., Anton, E., Teuteberg, F., & Dwivedi, Y. K. (2022). The role of the social and technical factors in creating business value from big data analytics: A meta-analysis. *Journal of Business Research, 153*, 128–149. doi:10.1016/j.jbusres.2022.08.028.

[7] Jones, G. L., & Qin, Q. (2022). Markov chain Monte Carlo in practice. *Annual Review of Statistics and Its Application, 9*, 557–578. doi:10.1146/annurev-statistics-040220-090158.

[8] Nobanee, H., Shanti, H., Aldhanhani, H., Alblooshi, A., & Alali, E. (2022). Big data and credit risk assessment: A bibliometric review, current streams, and directions for future research. *Cogent Economics & Finance, 10*(1),

DOI: 10.54254/2753-8818/2026.PU38792

2132638. doi:10.1080/23322039.2022.2132638.

[9] Li, C., Shah, N., Li, Z., et al. (2024). Modelling of wind and solar power output uncertainty in power systems based on historical data: A characterisation through deterministic parameters. *Journal of Cleaner Production, 484*, 144233.

[10] Ma, T., Liu, K., Qiu, Y., et al. (2025). Quantitative risk assessment of wellbore collapse of inclined wells in formations with anisotropic rock strengths. *Rock Mechanics and Rock Engineering, 58*(2), 1795–1818.

[11] Ryan, O., Haslbeck, J. M. B., & Waldorp, L. J. (2025). Non-stationarity in time-series analysis: Modeling stochastic and deterministic trends. *Multivariate Behavioral Research, 60*(3), 556–588.

[12] Brimacombe, M. (2025). P-values and laws of large numbers: Practical interpretation of significance in very large samples. *Population Medicine, 7*(April), 1–10.