From Uncertainty to Insight: The Fundamental Value of Probability Theory and Mathematical Statistics in the Era of Big Data

Yuqi Zhu
Zhejiang Hangzhou No.4 High School International School, Hangzhou, Zhejiang, China

Abstract: The defining feature of the big data era is not simply "massive data volume," but a fundamental shift in data dimensions: high volume, variety, velocity, and low value density. Against this background, the foundational value of probability theory and mathematical statistics has become increasingly prominent-the former provides a rigorous mathematical framework for uncertainty, while the latter offers systematic methods for extracting regularities from data. Based on the knowledge foundation of high school students, this paper systematically reviews the core concepts of probability theory and mathematical statistics and their deep connections with big data, analyzes their foundational value in fields such as prediction, decision-making, algorithms, quality control, and causal analysis, and illustrates them through three real-life cases: recommendation systems, epidemic forecasting, and learning analytics. The study shows that the relationship among the three can be aptly compared to "a building, its foundation, and its superstructure": probability theory is the theoretical foundation, mathematical statistics is the methodological bridge, and big data is the frontier of application. On this basis, the paper further discusses the implications for high school mathematics education: dispelling the myth that "mathematics is useless," cultivating statistical thinking and data awareness, and laying the groundwork for data literacy in the age of intelligence.

Keywords: Probability Theory; Mathematical Statistics; Big Data; Statistical Thinking; Data Literacy

1. Introduction

We live in an era surrounded by data. Short video platforms always push content you are interested in, music apps seem to "understand you" by recommending new songs, and epidemic trend charts have become a daily focus for many people-behind these phenomena lies the same set of mathematical tools: probability theory and mathematical statistics.

Yet many high school students ask: What do the probability formulas and statistical charts in textbooks have to do with the Douyin I browse or the NetEase Cloud Music I listen to every day? The misconception that "mathematics is useless" often stems from the cognitive gap between knowledge and reality. In fact, when you see the line "recommended based on your viewing history," a probability-based predictive model is running behind it; when you read a news headline that says "XXX new cases expected next week," statisticians are using mathematical methods to infer future uncertainties.

This paper aims to answer three core questions from a perspective accessible to high school students: What exactly do probability theory and mathematical statistics study? How do they underpin the technological applications of the big data era? And what does understanding them mean for our current mathematics learning?

2. Core Concepts: Building a Cognitive Bridge from the Classroom to the Frontier

2.1 Probability Theory: The Mathematical Language of Uncertainty

Life is full of uncertainty: Will it rain tomorrow? What is the probability of getting heads when flipping a coin? Probability theory is the branch of mathematics that studies the laws of random phenomena. Core concepts encountered at the high school level include:

· Random event: An event that may or may not occur, such as "the recommendation algorithm happens to correctly guess a video you like."

· Probability: A value between 0 and 1 that measures the likelihood of a random event occurring.

· Expectation: The "average level" of a random variable's values, such as "average daily video viewing time."

· Variance: An indicator measuring the degree of data dispersion, reflecting the magnitude of uncertainty.

The core idea of probability theory is: although a single random event cannot be predicted, stable statistical regularities emerge under a large number of repetitions. This is the very mathematical foundation that enables big data to "predict" the future.

2.2 Mathematical Statistics: Reading Regularities from Data

If probability theory is about "knowing the distribution, predicting the outcome," then mathematical statistics is about "knowing the outcome, inferring the distribution"-it is the science of extracting information from data. At the high school level, it mainly involves:

· Descriptive statistics: Summarizing data characteristics using the mean, median, variance, charts, etc., such as evaluating class performance through the average score.

· Inferential statistics: Using sample data to infer characteristics of a population, such as predicting the preferences of an entire city by surveying 1,000 people.

The distinction between the two can be understood this way: probability theory helps you calculate "roughly how many heads you would get when flipping a coin 100 times," whereas mathematical statistics, based on the actual result of "getting 53 heads out of 100 flips," works backward to infer "whether the coin is fair."

2.3 Big Data: The Application Scenario

If probability theory is the foundation and mathematical statistics is the bridge, then big data is the grand application scenario where the two converge. Big data is not merely about "massive amounts of data"; rather, it is characterized by the "4V" principle: Volume (large-scale data), Variety (diverse data types), Velocity (high-speed data generation and processing), and low Value density (a vast amount of data contains only a small proportion of valuable information). These four dimensions define the essential challenges of the big data era.

A central challenge in big data applications is noise. Real-world data is inevitably accompanied by measurement errors, missing values, duplicate records, and outliers. For example, in a recommendation system, a user may accidentally click on a video they have no interest in, which introduces "noise" into the behavioral data. Without proper noise handling, algorithms may learn spurious patterns and produce unreliable predictions. Therefore, data cleaning and preprocessing-grounded in statistical methods-are essential steps before any meaningful analysis can take place.

Furthermore, the design of algorithms in the big data context relies heavily on the 4V characteristics. For instance, the high Velocity of data streams requires algorithms to be computationally efficient and capable of online learning; the high Variety of data formats (text, images, click logs, etc.) necessitates the use of flexible statistical models that can handle heterogeneous inputs; and the low Value density demands that algorithms be equipped with effective feature selection and dimensionality reduction techniques. In essence, the 4V principle not only describes the nature of big data but also dictates the algorithmic strategies required to extract insights from it.

2.4 The Relationship Among the Three: The Analogy of a Building, Its Foundation, and Its Superstructure

The relationship among the three can be understood through an architectural analogy:

· Probability theory is the "theoretical foundation": it provides the mathematical language and axiomatic system for describing uncertainty. Without probability theory, it would be impossible to quantify "how likely the recommendation system is to guess correctly."

· Mathematical statistics is the "methodological bridge": it provides concrete methods for learning regularities from data. Without mathematical statistics, massive data would be nothing more than a chaotic pile of symbols.

· Big data is the "application superstructure": it is the engineering practice of probability theory and mathematical statistics in a "4V" environment. Together, they form a complete chain of "theory → method → application."

As Academician Liu Jun has pointed out: "Statistics is built on the cornerstone of probability theory, and through experimental design, data collection, and information extraction, it supports decision-making and prediction. Its core methods are the main driving force behind the current AI revolution." [5]The algorithms of the big data era are essentially extensions of probability theory and mathematical statistics in high-dimensional, large-scale settings.

3. Analysis of the Fundamental Value of Probability Theory and Mathematical Statistics

3.1 The Core Value of Probability Theory: Modeling Uncertainty

3.1.1 Providing a quantitative framework for uncertainty in big data

The "bigness" of big data does not eliminate uncertainty; rather, it amplifies it. Data missingness, measurement errors, and sampling bias are ubiquitous. Probability theory provides the mathematical language to quantify this uncertainty: probability distributions describe the possible values of data, confidence intervals express the credibility of estimates, and hypothesis testing judges whether observed differences are statistically significant.[2]

3.1.2 Supporting prediction and decision-making in big data

When you open NetEase Cloud Music, the generation of the "Daily Recommendations" playlist relies on a fundamental question: based on your past listening history, what is the probability that you will like a new song? The "Naive Bayes classifier" behind it is a direct application of Bayes' theorem. A recommendation system does not "read your mind"; it calculates the option that best matches your preferences.

3.1.3 Laying the underlying logic for machine learning algorithms

The core task of machine learning can be summarized as learning a mapping function from data. Whether it is supervised learning (predicting the future based on historical data) or unsupervised learning (discovering hidden structures in data), the underlying logic is inseparable from probabilistic models. For example, the Gaussian distribution (normal distribution) is the default assumption for many algorithms, and the "bias-variance tradeoff" is a key statistical concept for understanding a model's generalization ability.

3.2 The Core Value of Mathematical Statistics: From Data to Insight

3.2.1 Extracting core regularities from massive data

The "low value density" of big data means that massive amounts of data contain a great deal of useless noise. Mathematical statistics provides a complete set of tools for "refining" regularities: descriptive statistics (mean, variance, quantiles) can quickly characterize the overall picture of data; inferential statistics can make scientific judgments about a population based on samples.

3.2.2 Data cleaning and quality control

In the field of data science, there is a well-known saying: "Garbage in, garbage out." In the actual data analysis workflow, 80% of the time is spent on data cleaning. Statistical methods play a key role in this phase[2]:

· Outlier detection: Using the "3σ rule" (the probability that a data point falls beyond three standard deviations from the mean is less than 0.3%) to identify data entry errors.

· Missing value imputation: Filling in missing values with the mean, median, or model-predicted values.

· Data standardization: Transforming data of different units onto comparable scales.

3.2.3 Distinguishing correlation from causation

Big data often reminds us that "correlation does not imply causation," but the goal of statistics goes far beyond that. Correlation analysis (such as the correlation coefficient) can help discover patterns of association between variables; causal inference methods (such as randomized controlled trials, instrumental variable methods, etc.) attempt to answer the deeper question: "Does A cause B?"[6] This distinction is critically important in business decision-making and policy evaluation.

4. Real-Life Case Studies

4.1 Case One: The Secret of Recommendation Systems-Everyday Applications of Bayes' Theorem

Scenario: Douyin’s "For You" page, NetEase Cloud Music’s "Daily Recommendations," Bilibili’s homepage video feed.

Question: How do platforms filter out content you are interested in from massive amounts of information?

Statistical Principle: The core of a recommendation system lies in calculating a "conditional probability"-P(you like a video | your historical viewing records, likes, watch time, etc.). According to Bayes' theorem:

$P(A|B) = P(B|A) \times P(A) / P(B)$

Plain language explanation: The system comprehensively considers a video’s "baseline

popularity" and its "degree of match with your personal preferences" to update its judgment about your taste.
Value Embodied: Without probability theory, there would be no mathematical tool to quantify "uncertain preferences"; without mathematical statistics, it would be impossible to estimate model parameters from your behavioral data.[1]

4.2 Case Two: Trend Forecasting in Epidemic Transmission

Scenario: The "projected new case curve" and "basic reproduction number R0 and, in ongoing monitoring, the effective reproduction number Rt " familiar to the public since 2020.
Question: How do epidemiologists predict the number of cases for the coming week?
Statistical Principle: Time series analysis, regression models, confidence intervals. Historical data are used to fit a growth curve, and statistical models are used to extrapolate future trends, while also giving a range of uncertainty for the prediction (e.g., "an estimated 5,000–7,000 new cases expected").
Value Embodied: Forecasting is not "fortune-telling" but a scientific inference based on probability distributions. The "prediction interval" presented to the public is a direct application of the concept of confidence intervals in statistics.[6]

4.3 Case Three: Statistical Thinking in High School Learning

Scenario: Post-exam score analysis, class sampling surveys.
Question: Can the average score fully represent the level of a class? How should one design a sampling survey to understand students' sleep duration?
Statistical Principles:
· Variance and standard deviation: Two classes may have the same average score, but one has concentrated scores while the other is polarized-variance can quantify this difference.
· Stratified sampling: When surveying students' sleep habits, sampling proportionally by grade level (stratified sampling) is more representative than simple random sampling.

Value Embodied: Statistics is not just formulas; it is a way of observing the world-using data to speak, rather than judging by intuition.

5. Implications for High School Students' Learning

5.1 Dispelling the Myth That "Mathematics Is Useless"

When we see Bayes' theorem behind short video recommendations and regression models behind epidemic predictions, we should realize: every mathematical concept learned in high school is a stepping stone to cutting-edge technology.[3]Probability and statistics are not "armchair learning of no practical use," but the fundamental language of the intelligent age.

5.2 Cultivating Statistical Thinking and Data Awareness

The "General High School Mathematics Curriculum Standards" have already listed "data analysis" as one of the six core competencies.[4]This means the goal of mathematics learning is not only solving problems, but also developing the habit of "using data to discover, analyze, and solve problems." When encountering a problem, one can think: Can data help me make a decision? How credible is this conclusion?

5.3 Laying the Foundation for Future Learning and Career Development

Regardless of whether your future major is economics, medicine, psychology, or computer science, statistical thinking will become a core competency. As Progression of Statistical Competence in Academic and Career Pathways:
High School (Foundations of Probability and Statistics) → Undergraduate (Mathematical Statistics, Econometrics) → Graduate School (Machine Learning, Causal Inference) → Workplace (Data Analyst, Algorithm Engineer), from introductory statistics courses at the undergraduate level, to machine learning and causal inference at the graduate level, and then to data analyst positions in the workplace, probability and statistics form a continuous thread of knowledge. A solid foundation laid in high school will save a great deal of time in university studies.

6. Research Summary and Outlook

6.1 Core Conclusions

This study demonstrates that: (1) Probability theory and mathematical statistics are indispensable foundational disciplines in the era of big data; the former provides a quantitative framework for uncertainty, while the latter provides a methodological system for extracting regularities from data; (2) The relationship among the three can be summarized as "probability theory is the theoretical foundation, mathematical statistics is the methodological bridge, and big data is the frontier of application"; (3) Everyday applications such as recommendation systems and epidemic forecasting prove that probabilistic and statistical knowledge is not "ivory tower theory," but the underlying technology that shapes modern life.

6.2 Research Limitations

Constrained by the cognitive level of high school students, this study simplifies a large number of complex mathematical derivations and algorithmic details; the case selection is biased toward the consumer internet sector, with insufficient coverage of industrial big data, scientific big data, and other scenarios; the research depth is primarily a review, and no original data analysis experiments were conducted.

6.3 Future Outlook

With the deepening development of artificial intelligence, "explainable AI" has become a new focus of attention-this precisely requires the support of statistical causal inference methods. For high school students, the future direction of learning may be: while solidifying the foundation of probability and statistics, pay attention to data ethics (such as privacy protection and algorithmic fairness), learn basic data analysis tools (such as Python and Excel), and truly transform mathematical knowledge into problem-solving ability.

References

[1] Zhang Wei, Liu Ming. Analysis of the Application of Probability Theory and Mathematical Statistics in Data Mining [J]. Science & Technology Information, 2024(06).

[2] Gao Song. Expert Perspective: Quality Management and Probabilistic Thinking under Big Data [EB/OL]. InfinityQS, 2025-04-25.

[3] Wanshi. Programmer’s Mathematics (XIII): Mathematical Thinking in Data Processing and Analysis [EB/OL]. Huawei Cloud Community, 2025-12-21.

[4] Li Qiang. How to Cultivate Students' Data Analysis Literacy [N]. China Education Daily, 2025-09-22(05).

[5] Liu Jun. Thoughts on the Implementation of Artificial Intelligence-Viewing AI from Statistics [R]. Department of Statistics and Data Science, Tsinghua University, 2026.

[6] Chen Hao, Zhao Lei. The Full Process of Econometric Multi-Source Data Processing, Machine Learning Prediction, and Complex Causal Identification [EB/OL]. ScienceNet.cn, 2026-01-15.

[7] Han Jiawei, Micheline Kamber. Data Mining: Concepts and Techniques [M]. 3rd ed. Morgan Kaufmann, 2011.

[8] Pearl, Judea. Causality: Models, Reasoning, and Inference [M]. 2nd ed. Cambridge University Press, 2009.

[9] Wu Xiaoling. Research on the Application of Big Data Analysis Based on Probability Statistics [J]. Journal of Mathematics and Statistics, 2023(12).

[10] Breiman, Leo. Statistical Modeling: The Two Cultures [J]. Statistical Science, 2001, 16(3): 199-215.