Journal of International Education and Development
2025, VOL. 9, NO. 9, 193-196
DOI: 10.47297/wspiedWSP2516-250043.20250909

Research on the Teaching Reform of Probability and Statistics Courses in the Big Data Era

Yaxian Ma

School of Mathematics, North University of China, Taiyuan, Shanxi 030051, China

ABSTRACT

With the rapid penetration of big data technology, the demand for data analysis capabilities in various industries has gradually changed. Probability and statistics serve as the core course for cultivating data analysis thinking, yet the traditional teaching model lags behind in terms of content, methods, and evaluation systems. Therefore, this paper, based on the new requirements for talents' capabilities in the big data era, analyzes the existing problems in the current teaching of probability and statistics courses, such as the over-emphasis on theoretical derivation in content, the lack of interactivity in teaching methods, and the singleness of assessment methods. From three dimensions—updating teaching content, innovating teaching methods, and reconstructing the assessment system—it proposes reform strategies, including integrating big data cases from multiple fields, strengthening the teaching of Python/R tools, promoting the project-driven model, and constructing process-oriented evaluation. These strategies aim to enhance the adaptability of the course to big data applications, cultivate students' core competitiveness in solving practical problems, and provide references for talent cultivation under the background of emerging engineering and emerging liberal arts.

KEYWORDS

Big data era; Probability and statistics course; Teaching reform; Talent cultivation

1 Introduction

Probability and statistics is a basic course for multiple disciplines in higher education, including science and engineering, economics and management, and humanities. Its core value lies in cultivating students' thinking ability to extract rules from uncertainty and support decisions through data. With the in-depth application of big data technology in fields such as financial risk control, medical diagnosis, and online public opinion analysis, the scale of data has grown exponentially, data types have expanded from structured to unstructured, and analysis needs have shifted from descriptive statistics to predictive modeling and in-depth mining. This change has transformed the requirements for learners' capabilities. They are no longer limited to mastering theoretical knowledge such as classical probability models and hypothesis testing, but also need to possess comprehensive literacy in handling massive data, using professional tools for modeling, and interpreting analysis results [1]. However, the probability and statistics courses in most colleges and universities still follow the traditional teaching framework and fail to fully align with the technical characteristics and industry needs of the big data era. This leads to a disconnect between course teaching and practical application—students master theoretical formulas but struggle to deal with actual data problems. Against this backdrop, promoting the teaching reform of probability and statistics courses and reconstructing the course content and teaching model have become key issues to adapt to the development of the times and improve the quality of talent cultivation.

2 The necessity of teaching reform for probability and statistics courses in the big data era

2.1 New requirements for data analysis capabilities in the big data era

Traditional data analysis relies on small samples to infer the overall rules. The randomness and representativeness of sample selection directly affect the reliability of conclusions, and the analysis process is mostly limited to structured data. In the big data era, the cost of data acquisition has decreased, and various industries can collect full-scale data by virtue of technical means. The analysis logic has shifted from inferring the whole from samples to mining correlations from full data, without over-reliance on sample assumptions. This change in thinking requires learners to break out of the traditional statistical framework and understand new issues in massive data, such as noise processing and redundant information screening. At the same time, enterprises have more comprehensive demands for data analysis capabilities. Learners not only need to master theories such as probability distribution and regression analysis but also be proficient in using tools for data cleaning and visual presentation, reveal the business logic behind the data through modeling, and transform the analysis results into actionable decision-making suggestions. Single theoretical knowledge can no longer meet the job requirements.

2.2 Limitations of traditional probability and statistics course teaching

The traditional teaching of probability and statistics courses focuses on theoretical derivation. Classroom content mostly centers on formula proof and exercise calculation, and the selected examples are mostly idealized scenarios, such as classical probability problems like rolling dice and drawing balls, which are disconnected from the actual data characteristics in the big data environment [2]. Although students can master the application of formulas, they lack an understanding of the complexity of real data and often feel helpless when facing problems such as unstructured data and missing value processing. In addition, the integration of course teaching with computer science and domain knowledge is insufficient. There is no systematic introduction to mainstream data analysis tools such as Python and R, nor is the teaching content designed in combination with the business scenarios of different majors. This leads to a dilemma where students understand statistics but cannot apply it. Under this teaching model, there is an obvious gap between the course and cutting-edge technological applications, which fails to provide effective support for students' subsequent participation in practical projects.

2.3 Teaching reform is an inevitable choice for cultivating innovative talents

At present, the construction of emerging engineering and emerging liberal arts emphasizes interdisciplinary integration and practical innovation, requiring colleges and universities to break down traditional disciplinary barriers and cultivate talents with cross-boundary thinking and the ability to solve complex practical problems. As a bridge connecting mathematical theories and practical applications, the teaching quality of probability and statistics directly affects students' data analysis literacy and innovative ability. Under the traditional teaching model, students passively accept knowledge, lack opportunities for independent thinking and teamwork, and their innovative thinking is difficult to stimulate. Through teaching reform, adjusting the course content and teaching methods can transform abstract statistical theories into tools for solving practical problems, guiding students to shift from "learning knowledge" to "applying knowledge".

3 Analysis of the current teaching situation of probability and statistics courses

3.1 Teaching content

The current teaching content of probability and statistics courses still focuses on traditional mathematical knowledge. Most textbook chapters revolve around theoretical modules such as random events and probability, random variables and their distributions, parameter estimation, and hypothesis testing. The emphasis on formula derivation is far greater than that on practical application. The selected cases are mostly simplified ideal scenarios, such as product qualification rate calculation and mean estimation of height and weight data. These cases have small data volume and simple structure, which are significantly different from the characteristics of massive, multi-source, and unstructured data in the big data environment. It is difficult for students to understand the application logic of statistical methods in practical scenarios through these cases. At the same time, the course has a serious lack of introduction to modern data processing tools. Most classes only briefly mention the statistical functions of Excel at the end of the course, and there is no systematic explanation of the core libraries of professional tools such as Python and R (e.g., Pandas for data processing and Matplotlib for visualization). This results in students lacking the ability to use tools to process real data, leading to a disconnect between course learning and practical needs.

3.2 Teaching methods

Most colleges and universities still adopt the one-way teaching model of "teacher lectures + students listen" for probability and statistics courses. Classroom time is mainly used for theoretical explanation and example calculation, and students lack opportunities for active participation. Under this model, teachers dominate the classroom rhythm, and students passively accept knowledge, making it difficult for them to think in depth about difficult issues. Classroom interaction is mostly limited to simple questions like "Do you understand?", lacking in-depth discussions on analysis ideas and method selection. The weakness of practical teaching is another prominent problem. The practical part of most courses only involves completing statistical exercises after class or conducting simple Excel data calculations in computer labs, lacking complete analysis projects based on real data. Even if some courses set up experimental links, they are mostly verification experiments where students only need to operate according to preset steps without independently designing analysis plans, which fails to cultivate their ability to solve problems independently [3]. The low correlation between course projects and actual problems makes it difficult to stimulate students' learning interest and practical enthusiasm.

3.3 Assessment and evaluation

The assessment methods of probability and statistics courses are generally single, with the final closed-book exam as the main basis for evaluation. The exam content mostly focuses on theoretical knowledge and formula application, such as probability calculation, distribution function solution, and hypothesis testing step writing, accounting for a too high proportion of the total score. This assessment method overemphasizes students' memory and mastery of theoretical knowledge while ignoring the evaluation of data processing processes, innovative thinking, and teamwork skills. The usual grades are mostly composed of after-class homework and attendance. The homework content is mainly based on textbook exercises, lacking the assessment of students' analytical ability and tool application ability. To cope with the final exam, students often adopt the method of rote memorization of formulas and extensive problem-solving. Although they can achieve high scores, they cannot flexibly apply statistical methods to practical problems. The assessment results cannot fully reflect students' data analysis literacy, nor can they provide effective feedback for teaching improvement, which is not conducive to the improvement of talent cultivation quality.

4 Strategies for the teaching reform of probability and statistics courses in the big data era

4.1 Promote the update and integration of teaching content and cases

4.1.1 Integrate big data cases

Break the limitations of traditional idealized examples and introduce real big data cases from multiple fields. For example, in the chapter of "Probability Distribution", use user order data from online shopping platforms to explain the application of Poisson distribution in traffic prediction; in the "Regression Analysis" module, use financial credit data to enable students to analyze the impact of customer factors on default risks. At the same time, introduce data from fields such as biomedicine, transportation, and social media to meet the needs of different majors. Case teaching should present the entire process of data acquisition, cleaning, analysis, and conclusion output, enabling students to understand the complexity of real data and cultivate their ability to solve practical data problems.

4.1.2 Strengthen software tool teaching

Take Python or R language as the core teaching tool and adopt a synchronous model of theory and practical operation. For instance, when explaining "descriptive statistics of data", first introduce the theoretical concepts, then use Python's Pandas library to demonstrate the calculation of real datasets and visualize the results through the Matplotlib library. Or when explaining "hypothesis testing", first elaborate on the principles, then guide students to complete the test using the Scikit-learn library. Compile supporting tool application manuals and set phased tasks, such as "analyzing e-commerce user consumption data with R language", requiring students to independently complete data processing and modeling to exercise their ability to use big data tools [4].

4.1.3 Add interdisciplinary content

On the basis of traditional statistical knowledge, add modules such as basic machine learning, data visualization, and big data ethics. After the chapter of "Probability Models", briefly introduce the Naive Bayes algorithm and its application in spam classification. After the "Data Distribution" module, explain visualization methods such as heat maps and word clouds. In the later stage of the course, set up a special topic on big data ethics to discuss issues such as data privacy protection. The interdisciplinary content aims at understanding principles and mastering applications, avoiding in-depth technical details, and cultivating students' interdisciplinary thinking.

4.2 Innovate diversified teaching methods and models

4.2.1 Promote project-driven teaching

Design 2-3 small research projects according to the progress of the semester. The themes should combine students' majors and big data scenarios. For example, design an e-commerce user consumption analysis project for economics and management majors, and a public opinion data mining project for computer majors. The projects are carried out by teams of 3-4 students, covering the processes of data acquisition, preprocessing, analysis, visualization, and report output. Teachers regularly organize team reports, provide feedback and technical guidance, and collect analysis reports and codes after the completion of the projects to cultivate students' teamwork, problem-solving, and innovative thinking.

4.2.2 Adopt a blended teaching model

The blended teaching model combines the flexibility of online learning with the interactivity of offline classrooms, which can effectively improve teaching efficiency and quality. The course can rely on online platforms such as MOOC and SPOC to create online course resources for theoretical knowledge modules (such as basic probability concepts, random variable distribution, and parameter estimation principles), including video explanations, courseware, and practice questions. Students are required to independently learn the online content before class, complete basic practice questions, and master core theoretical knowledge. Offline classrooms focus on difficult problem answering, case discussions, and practical operation guidance. For example, for the problem of "application of Bayes' theorem" that students generally find confusing in online learning, conduct in-depth explanations through classroom cases (such as probability calculation in disease diagnosis); organize students to discuss the application limitations of the theoretical knowledge learned online in big data scenarios, such as "whether small-sample statistical inference still has value in the full-data environment"; carry out practical teaching in computer labs, guide students to use Python/R tools to complete data processing and analysis tasks, and promptly solve problems encountered by students in code writing and tool use. The blended teaching model can adjust the focus of offline teaching based on students' online learning data (such as viewing duration and exercise accuracy rate) to achieve personalized teaching. At the same time, it saves time for theoretical explanation and reserves more space for practical links and in-depth discussions.

4.2.3 Conduct seminar-style and case-based teaching

Seminar-style and case-based teaching can cultivate students' critical thinking and data interpretation abilities, breaking the limitations of one-way teaching. In class, group presentations and collective discussions are adopted: each group shares their understanding, while other groups raise questions and provide supplementary insights. Teachers guide students to think in depth, using prompts such as "Will the conclusion change if a different data processing method is adopted?" Additionally, special seminars on hot topics can be organized, such as "The Value of Statistical Inference in the Big Data Era," to encourage students to express their views and foster their critical thinking and academic expression abilities.

4.3 Construct a process-oriented assessment and evaluation system

4.3.1 Implement diversified assessment

Construct a "process-oriented evaluation + summative evaluation" system, reduce the proportion of the final exam score, and compose the rest of the score from usual homework, experimental reports, project results, and classroom performance. Design practical tasks for usual homework, require experimental reports to record the complete analysis process, score project results in combination with presentation and defense, and evaluate classroom performance based on the degree of participation in discussions. This comprehensive coverage of the learning process can objectively reflect students' data analysis literacy.

4.3.2 Focus on competence evaluation

The assessment shifts from knowledge mastery to competence improvement. The evaluation of projects and experimental reports is carried out from four dimensions: clarity of thinking, applicability of methods, standardization of code, and ability to explain results. For example, if a student's method is not the optimal one but has clear logic, standardized code, and reasonable explanation, a high score is still given; if the result is correct but lacks logical explanation, the score is reduced. Reduce the number of theoretical memory questions in the final exam and increase case analysis and application questions to test the ability to apply knowledge, guiding students to shift from exam-oriented learning to comprehensive competence improvement.

5 Conclusion

The big data era puts forward new requirements for talents' data analysis capabilities. As the core carrier for cultivating this capability, the teaching reform of probability and statistics courses is imperative. The current backwardness of the course in content, methods, and assessment system leads to a disconnect between teaching effects and practical needs, failing to fully support the talent cultivation goal. By promoting the integration of teaching content with big data cases, professional tools, and interdisciplinary knowledge, innovating diversified teaching methods such as project-driven, blended, and seminar-style teaching, and constructing a process and competence-oriented assessment and evaluation system, the adaptability of the course to the times and its practical value can be effectively enhanced.

About the Author

Yaxian Ma (1994-11-), female, Ph.D., Lecturer. Research direction: Partial Differential Equations.

References

[1] Xie G , Jia Y .Teaching Reform Strategy of Ideological and Political Course Based on Big Data-Taking "Digital Electronic Technology" as an Example[J]. Journal of Physics: Conference Series, 2021, 1852(3):032003 6pp).

[2] Huang T , Dong J , Min L .Teaching Reform and Practice of Statistics Courses in Big Data Management and Applications Major in the Context of New Quality Productivity[J]. Journal of Contemporary Educational Research, 2025, 9(2):23-31.

[3] Cheng S I .Discussion on the Application of Case Teaching Method in Civil and Commercial Law Teaching in Higher Vocational Colleges[J]. Journal of International Education and Development, 2022.

[4] Liu H , Sun G .Research on the Application of Big Data Technology in the Integration of Enterprise Business and Finance[J]. Big Data Journal (English Edition), 2021.