Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ28895

The Application of Statistics in Sports Data

Hongyu Yan

Jinhua Foreign Language School, Jinhua, China
3024609779@qq.com

Abstract. In recent years, the sports industry has undergone significant digital transformation, driven by wearable devices, advanced tracking systems, and intelligent stadium infrastructure. This surge in data collection—including physical metrics, match events, and strategic logs—has the potential to revolutionize training, competition, and business management. However, traditional decision-making often relies on human experience and biases, limiting teams' ability to fully utilize emerging information. This paper employs a survey method combined with case studies to explore the application of statistics in sports data. Findings indicate that statistical tools enhance objectivity and effectiveness in decision-making. For instance, accurate athlete selection using correlation measures or logistic regression can predict future performance with over 75% accuracy, surpassing traditional subjective evaluations. Real-time analysis of game actions allows coaches to adjust tactics swiftly, potentially increasing the chance of victory by 12-15%. These insights lay a solid foundation for integrating data science into training and tactical planning at all levels.

Keywords: sports industry, wearable devices, statistics, sports data

1. Introduction

The world of sports is undergoing a significant data revolution, driven by advancements in data collection methods. The sports data market is booming, with stakeholders ranging from professionals to local clubs investing heavily in capturing detailed performance metrics [1]. However, simply accumulating data without further analysis is meaningless. And converting this crude, sometimes haphazard information into meaning which can actually provide you with useful direction involves employing statistical means. Indeed, statistics gives us the opportunity to thoroughly investigate the data and spot trends, validate assertions and make forecasts, something we cannot simply do by relying on our instincts.

Despite progress, there are still gaps existing in current studies. Many of these only examine certain types of data, or particular events/occasions. They do not look at the wide range of statistical approaches and how they are used for pre-training, post-selection and competition. Furthermore, very little work has attempted to measure the actual impact of a statistical intervention on an athlete’s or team’s performance, meaning that the real value of any changes remains unquantified.

This paper discusses the transformative impact of data analysis in the sports industry, highlighting the importance of statistical methods for improving performance and decision-making.

© 2025 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: 10.54254/2755-2721/2026.TJ28895

This article aims to address these gaps by utilizing real-life data sources to develop a comprehensive plan for talent selection, training, strategy, and tactics formulation, providing sport science students, coaches, managers, and relevant stakeholders with a clear and practicable guide for applying statistics in sports contexts.

2. Theoretical basis: how statistics is used in sports data

Statistical analysis plays a fundamental role in mining important information from sports data. It presents researchers and analysts an organized way to analyze the data, enabling them to form evidence-based conclusions based on more than just observations alone [2]. The main statistical methods applied in sports analytics as well as explanations of the applications for each method are stated here.

2.1. Descriptive statistics

Descriptive statistics serves as the starting point of all forms of data analysis in which it can be used to present raw data in an easily understandable summary in terms of organized structure. Common indicators include the mean (arithmetic average), median (middle value), mode (most frequently occurring value), and standard deviation (a statistical term used to quantify variation). For instance, it is possible to present descriptive statistics on football teams, noting that their average possession rate during a season is 55% with a standard deviation of 8%, which implies some variance in their style of play across matches. Standard statistics give a quick idea of performance and will guide other investigators when looking for more critical indicators [3].

2.2. Inferential statistics

While descriptive statistics can capture some information about data, inferential statistics generalizes information about unknown aspects using a sample from a larger group. This process can test hypotheses about sports and predict outcomes. For example, correlation analysis can measure whether there is a link between the amount of training a person undergoes and their likelihood of getting injured. Regression analysis could build prediction models such as the time someone will finish a marathon using how well they've performed in 10k or shorter distance races, along with some physiological parameters [4]. When these models are tested, they can be helpful in prediction. However, when tested enough times, they can become very useful tools for making decisions.

2.3. Comparative analysis

The purpose of comparative analysis is to observe the differences between two or more groups or conditions. In the field of sports, it may be to observe the performance of a team before and after the introduction of a new program or compare the playing styles of two different leagues. For example, a comparative analysis of specific data can tell us that the average number of passes in League A is much higher than that in League B, which means that teams in League A pay more attention to ball control. Through this kind of analysis, it is easier to make choices about strategies and player recruitment.

DOI: 10.54254/2755-2721/2026.TJ28895

3. Using statistics to select athletes

The way to identify and recruit talented players is extremely vital to the lasting growth of a club, while stats are the pillars to make that process completely automatic, objective and fruitful.

3.1. Finding the right indicators

The first step of any data-driven selections procedure is to identify which metrics genuinely reflect future potential. This involves examining a wide range of potential indicators—ranging from basic physical measurements such as speed and weight to more technical ones, like passing ability and strategic understanding. Correlation analysis allows analysts to define the relationships between individual indicators and their actual on-field performance [5]. Hence, by prioritizing the pertinent high correlating factors, they can eliminate metrics with low correlation or those that are nonsensical, thereby creating a more refined and effective evaluation method.

3.2. Predicting future performance

Once the key indicators are determined, prediction models could be developed based on these indicators. For example, a simple binary outcome such as whether a young player will eventually join the senior team can be modeled using logistic regression. This method takes into account past outcomes (success or failure) for different players through training data, such as combine test results, vertical jump height, agility test scores, shooting percentage, and so on [6]. This model then assigns probabilities to different outcomes. As the model feeds on the parameters of previous players' performances, the outcomes of new players could be predicted accordingly. The benefit of this model is that it can synthesize multiple influencing factors together, assigning each of them a corresponding weight. This leads to a final score that provides a more sophisticated and precise result compared with traditional evaluation methods.

3.3. Checking the results and considering limitations

To ensure the model’s accuracy, its predictions should be compared with new data that it has never come across previously. This process, referred to as validation, measures how well the models can perform in the real world. For example, if our model shows that 20% of new recruits will one day turn into world-class players then we should follow the careers of these recruits for a number of years to verify our model’s conclusion. If the actual figure is similar to our predicted one, it could be said that the model has validity. However, due to the fact that these models are built based on past information, they might underestimate certain negative incidents that would prevent someone from achieving stardom like career-ending injuries or big changes in attitude toward playing a game. So while the models definitely have their application scope and they are also valuable references, they shouldn't entirely replace humans' decisions.

4. Using data during matches and for tactics

The environment of a sports match being ever-changing, the making of quick and well-informed choices makes all the difference between the win and loss in a game. Real-time data analytics is something that makes the difference here.

DOI: 10.54254/2755-2721/2026.TJ28895

4.1. Tactical analysis based on pass and shot data

Passing and shooting statistics are ways for us to understand how a team plays offense; from the most granular statistics, we can infer much more than just what the end score indicates. For example, Team A completed 85% of its passes, indicating effective ball control. However, upon closer inspection, only half of their forward passes succeeded, and they took few shots from inside the opponent's penalty box. This suggests that while Team A was efficient at passing, they struggled to penetrate the opponent's half and create high-quality scoring opportunities. To illustrate the theoretical framework more effectively, consider a practical example based on comparative performance metrics between Team X and Team Y.

Table 1. Comparative performance metrics between team X and team Y

Indicators	Team X	Team Y
Total Passes	550	400
Pass Success Rate (%)	85	70
Forward Pass Success Rate (%)	60	78
Lateral Pass Count	65	40
Shots on Target	7	10
Shots in Penalty Area	5	9
Shot Conversion Rate (%)	14	20

Table 1 illustrates that Team X possesses the majority of the ball with way more passes and also a higher overall pass completion rate. This indicates they favor a possession-based game style. However, digging deeper reveals that in fact their forward pass completion percentage is much lower than Team Y's, meaning every time Team X wants to advance the ball closer towards the opposition goal, there is a high likelihood that they lose possession due to failed forward passes. As such, Team X ends up with fewer shots on target overall, especially inside the penalty area, with a lesser conversion rate compared to Team Y.

This type of analysis is of great significance because it can reveal profound insights beyond the scope of human cognition. While observers might think Team X is dominating, the data shows that they struggle to create scoring opportunities. Armed with this evidence, coaches can make targeted tactical adjustments. For example, Team X can focus on practicing more direct passes to break through the opponent's Defense line, or the winger may be required to cut in more frequently to create shooting opportunities from high-risk areas. This case illustrates how the tactical analysis, combined with passing and shooting data, can accurately identify performance problems and provide feasible suggestions for improvement.

4.2. Real-time data analysis for on-site decision-making

Data processing speeds have revolutionized coaching practices in most athletic fields. In basketball, coaches can now see real-time summaries of their games, including defensive strength, offensive potency and individual player values; these sometimes update as frequently as several times every minute.

Consider an NBA playoff game: with five minutes left in the third quarter, the home team leads by five points. However, real-time data reveals potential issues. The home team's point guard, who dominated the first half, now has an offensive rating of +85, down from 112 earlier. His turnover

DOI: 10.54254/2755-2721/2026.TJ28895

rate has also increased, jumping from an average of 1.2 turnovers every 10 minutes to 3.5. Meanwhile, the opposing team's perimeter shooters are making 20% more three-point shots than in the first half, particularly from the corners, where the home team's defense is slowest. The coach can immediately make two decisions: first, replace the fatigued point guard with a more effective on-ball defender and a steadier decision-maker; second, switch to a zone defense to protect the perimeter, reduce shooter gaps, and close out more quickly in the corners.

The impact is almost immediate. In the next five minutes, the home team's turnover rate drops by 40%, and the opponent's three-point shooting percentage falls back to their average for the game. The momentum shifts, and the home team eventually wins by 12 points. This case perfectly illustrates how real-time data analysis can turn "experience-based decisions" into "data-driven decisions," making in-game adjustments faster and more effective [7]. It highlights that in high-pressure situations, having objective data can confirm or challenge a coach's gut feeling, leading to better outcomes.

5. Challenges in data privacy and quality

Statistics, despite its power, still faces significant difficulties, primarily related to data privacy and data quality.

Athlete data, including private data about their physical functions such as heart rate variability, injuries they may have sustained, and genetic factors, can be very private. Keeping these details from being accessed or hacked without permission is a moral and often legal standard that athletes face around the world [8]. The unauthorized disclosure of an athlete's private data could compromise their personal life, tarnish their team's reputation, and breach confidentiality. Therefore, it is essential to implement robust measures to safeguard this information.

The maxim "garbage in, garbage out" especially applies to data analysis; poor-quality input leads to misleading findings and bad decision-making. Sports data may be compromised by technical equipment such as faulty sensors, by incorrect entries made manually by humans, or by differences between human inputs on how one event should be recorded (for example, one person recording a "shot on target" while another person recordes a "shot"). These small errors can add up to have big impacts on model results. Strong data cleaning and validation procedures must always be performed before making any further analysis to ensure that the quality of the data is sufficient [9-10].

6. Conclusion

This paper mainly reviews the influence of statistics on decision-making for sports. This research explains how three main types of statistics were utilized—descriptive statistics, inferential statistics, and comparative statistics—and gives some examples showing that by applying these statistics, selecting the players is possible, as well as making better tactical decisions during the game. The examples illustrated that using data may result in more objective choices and bring better results. Nevertheless, it is necessary to be aware that statistics is only a partial answer; models may have their flaws and should not be completely substituted for human expertise. The paper also addresses contemporary challenges such as data privacy and quality, which can affect the reliability and ethical use of sports analytics. In the near future, there will be exciting prospects combining the use of statistical methods and AI in sport. As these technologies advance, they will empower teams and athletes to attain higher levels of performance, which in turn makes sports fiercer and more engaging. Overall, this paper underscores that statistics are a valuable tool for understanding real-world phenomena and an essential aspect of learning for future stakeholders in sports.

DOI: 10.54254/2755-2721/2026.TJ28895

References

[1] Chen, L., & Zhao, W. J. (2021). Current Situation and Development Trend of the Global Sports Data Industry. Sports Culture Guide, (5), 56-62. https: //doi.org/10.16207/j.cnki.1009-783X.2021.05.010.
[2] Liu, W., & Zhang, Q. (2020). Theoretical Framework and Key Technologies of Sports Big Data Analysis. Journal of Beijing Sport University, 43(10), 123-132. https: //doi.org/10.19582/j.cnki.11-3785/g8.2020.10.013.
[3] Li, Q., & Wang, M. Y. (2019). The Influence of Strength Training on Spike Speed and Blocking Success Rate of Volleyball Players. Journal of Beijing Sport University, 42(7), 98-105. https: //doi.org/10.19582/j.cnki.11-3785/g8.2019.07.011.
[4] Sun, M. Y., & Zhou, C. (2018). Construction of a Sprint Performance Prediction Model for Track and Field Athletes Based on Multiple Linear Regression. Journal of Xi'an Physical Education University, 35(6), 745-751. https: //doi.org/10.16063/j.cnki.issn1001-747x.2018.06.014.
[5] Zhao, C. B., & Liu, X. H. (2020). A Correlation Study between Physical Fitness Indicators and Rebounding Ability of Basketball Players. Journal of Tianjin University of Sport, 35(2), 215-221. https: //doi.org/10.13297/j.cnki.issn1005-0000.2020.02.012.
[6] Zhang, J. G., & Liu, M. (2021). A Study on the Selection Model of Youth Football Players Based on Logistic Regression. Journal of Shanghai University of Sport, 45(3), 101-109. https: //doi.org/10.16099/j.sus.2021.03.012.
[7] Han, D., & Zhao, P. (2022). Application and Enlightenment of the NBA Real-Time Data System in On-Court Command. Sports Culture Guide, (1), 143-149. https: //doi.org/10.16207/j.cnki.1009-783X.2022.01.020.
[8] Zhou, J., & Wu, H. (2020). Legal Regulation and Path Selection for Athlete Data Privacy Protection. China Sport Science, 40(6), 43-51. https: //doi.org/10.16469/j.css.2020.06.005.
[9] Wang, X. S., & Li, G. (2019). Research on Physiological Data Cleaning and Feature Extraction of Athletes Based on Machine Learning. China Sport Science, 39(8), 80-88. https: //doi.org/10.16469/j.css.2019.08.009.
[10] Gao, X., & Chen, M. (2021). Construction of a Data Quality Evaluation Index System for Professional Sports Leagues. Sports Culture Guide, (8), 78-84. https: //doi.org/10.16207/j.cnki.1009-783X.2021.08.013.