InSight Bulletin: A Multidisciplinary Interlink International Research Journal
Peer Reviewed International, Open Access Journal.
ISSN : 3065-7857 / Website: https://ibrj.us / Volume-3, Issue-1 / January - 2026

Original Article

A Statistical Framework for Detecting Algorithmic Bias in Machine Learning Models Using Hypothesis Testing

Vidya Bhatlavande
ATSS CBSCA

Manuscript ID:
IBMIIRJ -2026-030135

Submitted: 09 Dec. 2025

Revised: 13 Dec. 2025

Accepted: 08 Jan. 2026

Published: 31 Jan. 2026

ISSN: 3065-7857

Volume-3

Issue-1

Pp. 182-185

January 2026

Correspondence Address:
Vidya Bhatlavande
ATSS CBSCA
Email:
vidyabhatlavande@atsscollege.edu.in

Quick Response Code:

Web. https://ibrj.us

DOI: 10.5281/zenodo.18955041

DOI Link:
https://doi.org/10.5281/zenodo.18955041

Creative Commons

Abstract

As machine learning (ML) systems become deeply embedded in contemporary decision-making processes, concerns regarding algorithmic bias have attracted increasing scholarly and societal attention. Automated models are now widely used in high-impact domains including recruitment, credit approval, education, and the criminal justice system, where discriminatory outcomes may arise, where biased outcomes may reinforce existing inequalities. Consequently, fairness has become a fundamental requirement rather than an optional design goal. Although many fairness-aware ML approaches rely on descriptive performance metrics such as accuracy, precision, recall, or selection rates, these measures alone are insufficient to determine whether observed disparities between demographic groups are statistically meaningful or merely due to random variation.

This paper proposes a simple yet rigorous statistical hypothesis testing framework for detecting algorithmic bias by formally comparing model outcomes across protected groups. The framework employs classical statistical tools, including the two-proportion z-test and the chi-square test of independence, to evaluate group-level differences in decision outcomes. A small synthetic dataset is used to demonstrate the proposed methodology in a transparent and interpretable manner. The results illustrate that statistically significant bias can be detected even when overall model performance appears balanced. The study emphasizes the importance of incorporating uncertainty and statistical significance into fairness assessments

Keywords: *Machine learning (ML) models are progressively utilized in diverse real-world applications*

Introduction

Machine learning (ML) models ensuring fairness and ethical decision-making has become essential to automate or support decision-making processes that directly affect individuals and communities. Applications such as résumé screening, credit approval, medical diagnosis, and student admissions are often promoted for their potential to improve efficiency, scalability, and consistency. Nevertheless, a growing body of empirical evidence demonstrates that these systems can unintentionally reproduce or amplify existing social and institutional biases present in historical data.

A prevalent assumption in applied machine learning is high predictive accuracy is commonly viewed as an indicator of fairness, despite the absence of guarantees regarding equitable treatment across demographic groups. However, this assumption is frequently violated in practice. Models may exhibit strong aggregate performance while systematically disadvantaging specific demographic groups. Such group-level disparities are often obscured when model evaluation relies solely on overall performance metrics, thereby limiting the ability to identify and mitigate biased outcomes.

In this work, I overcome this limitation by introducing traditional statistical hypothesis testing as a clear and practical solution for detecting algorithmic bias. Rather than relying exclusively on descriptive fairness measures, hypothesis testing provides a principled mechanism to evaluate if the observed outcome differences between population groups are statistically significant or attributable to random variation. The primary contribution of this paper is a statistically grounded and computationally simple framework that can be readily integrated into standard ML evaluation pipelines without requiring modifications to existing learning algorithms.

Background and Related Work

1. Algorithmic Bias and Fairness

Algorithmic bias refers to systematic and unfair outcomes produced by automated decision systems, particularly when they disadvantage protected or vulnerable groups. Barocas and Selbst (2016) show that bias can emerge even when sensitive attributes are excluded from model inputs, as other variables may act as proxies for protected characteristics.

Creative Commons (CC BY-NC-SA 4.0)
This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Public License, which allows others to remix, tweak, and build upon the work noncommercially, as long as appropriate credit is given and the new creations ae licensed under the idential terms.

How to cite this article:
Bhatlavande, V. (2026). A Statistical Framework for Detecting Algorithmic Bias in Machine Learning Models Using Hypothesis Testing. InSight Bulletin: A Multidisciplinary Interlink International Research Journal, 3(1), 182–185. https://doi.org/10.5281/zenodo.18955041

To address this issue, several fairness definitions have been proposed in the literature, including:

Demographic Parity, which focuses on equal selection rates Equalized Odds and Equal Opportunity evaluate fairness by comparing error rates between groups, whereas Calibration requires predictive scores to be equally meaningful across demographics (Hardt et al., 2016) demonstrate that these fairness notions are often incompatible, meaning that improving fairness under one definition may worsen it under another. As a result, fairness evaluation must be context-sensitive and carefully justified.

2. **Limitations of Existing Approaches**

In many applied studies, fairness is assessed using point estimates, such as differences in selection rates or error rates. While informative, these measures have important limitations:

- They ignore sampling variability
- They do not quantify uncertainty
- They may lead to misleading conclusions, especially for small or moderate datasets

Recent research highlights the need for greater statistical rigor in fairness assessments (Kallus et al., 2020). Hypothesis testing directly addresses this gap by enabling formal significance testing of observed disparities.

Methodology

1. **Problem Setting**

We consider a binary classification model that produces one of two outcomes:

- **Positive prediction** (e.g., approve, accept, select)
- **Negative prediction** (e.g., reject, deny)

The population is divided into two demographic groups:

- **Group A**
- **Group B**

The central question is whether the model's decisions are independent of group membership, or whether systematic differences exist between groups.

2. Statistical Hypothesis Tests

1. Two-Proportion Z-Test

To compare the rate of positive predictions between two demographic groups, we employ the two-proportion z-test. This test evaluates whether the observed difference in selection rates between Group A and Group B is statistically significant. Let pA and pB denote the true proportions of positive predictions (selection rates) for Group A and Group B, respectively. The hypotheses are formulated as:

H0: pA = pB (No difference in positive prediction rates between the two groups)

H1: pA ≠ pB (A difference exists in positive prediction rates between the two groups)

Suppose nA and nB represent the total number of instances in Group A and Group B, and

Let nA and nB represent the total number of observations in Groups A and B, respectively, while xA and xB denote the counts of positive predictions within each group. The corresponding sample proportions are defined as pA=xA/nA and pB=xB/nB,

where xA and xB indicate the number of favorable outcomes for Groups A and B, and nA and nB refer to their respective sample sizes. Under the null hypothesis, the two population proportions are assumed to be equal, and a pooled estimator of the common proportion is used:

$$\hat{p} = \frac{x_A + x_B}{n_A + n_B}$$

The test statistic is then computed as:

$$Z = \frac{\hat{p}_A + \hat{p}_B}{\sqrt{\hat{p}\,(1-\hat{p})\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}}$$

Under the null hypothesis and assuming sufficiently large sample sizes, the test statistic approximately follows a standard normal distribution. The corresponding two-sided p-value is obtained by comparing the absolute value of z to the standard normal distribution.

If the p-value is less than a predefined significance level α (commonly 0.05), the null hypothesis is rejected.

2. **Chi-Square Test of Independence**

The chi-square test examines whether group membership and model outcomes are statistically independent.

Null Hypothesis (H_0):

Group membership and model decision are independent.

Alternative Hypothesis (H_1):

Group membership and model decision are associated.

$$\chi^2 = \sum_i \sum_j \frac{(Oij - Eij)^2}{Eij}$$

A significant result indicates that predictions depend on group membership, suggesting potential bias.

Assumptions

The proposed framework relies on the following assumptions:

- Observations are independent
- Sample sizes are sufficiently large for asymptotic approximations
- Outcomes are binary and groups are mutually exclusive

Experimental Illustration Using a Synthetic Dataset

1. **Dataset Description**

Group	Positive Predictions	Negative Predictions	Total
A	40	60	100
B	20	80	100

The dataset is synthetic and intentionally simplified to clearly demonstrate the bias detection process. Both groups have equal sample sizes, ensuring that observed differences arise from model outcomes rather than group size imbalance.

Results

Descriptive Analysis

- **Selection Rate (Group A):** 40%
- **Selection Rate (Group B):** 20%

Although the model's overall accuracy is similar for both groups, Group A receives positive predictions at **twice the rate** of Group B.

Hypothesis Testing Results

- **Two-Proportion Z-Test:**
  - The test statistic indicates a statistically significant difference
  - p-value < 0.01
- **Chi-Square Test of Independence:**

The chi-square statistic is significant at the 1% level ($p < 0.01$), and thus, the null hypothesis is rejected by both tests Both tests reject the null hypothesis, providing strong statistical evidence of algorithmic bias.

Discussion

Key Insights

The analysis highlights several important insights:

- Aggregate accuracy metrics can mask group-level disparities
- Statistically significant bias may exist even when performance appears balanced
- Hypothesis testing offers objective and interpretable evidence for fairness auditing

Practical Implications

In real-world deployments, such biases can:

- Systematically disadvantage specific groups
- Increase legal and regulatory risks
- Undermine public trust in AI systems

Therefore, organizations deploying ML models in sensitive domains should adopt **routine statistical bias audits** as part of responsible AI practices.

Limitations

This study has several limitations:

- The use of synthetic rather than real-world data
- A focus on binary classification and two demographic groups

These limitations highlight opportunities for further research rather than weaknesses of the proposed framework.

Conclusion and Future Work

This paper demonstrates that **classical statistical** hypothesis testing provides a clear, accessible, and effective approach to detecting algorithmic bias in machine learning models. By incorporating significance testing into fairness evaluations, practitioners can avoid misleading conclusions based solely on descriptive metrics.

Future research directions include:

- Studying intersectional bias involving multiple protected attributes
- Extending the framework to regression and multi-class settings
- Integrating statistical testing with causal inference and bias mitigation methods

Overall, the proposed framework strengthens the connection between traditional statistical reasoning and modern ethical AI practices.

Acknowledgement

I would like to express my sincere gratitude to all those who supported and guided me in the completion of this project titled "A Statistical Framework for Detecting Algorithmic Bias in Machine Learning Models Using Hypothesis Testing."

First and foremost, I am deeply thankful to my project supervisor/mentor for their invaluable guidance, expert insights, and continuous encouragement throughout this research. Their support helped me develop a strong understanding of statistical methodologies, hypothesis testing techniques, and fairness evaluation in machine learning systems.

Financial support and sponsorship

Nil.

Conflicts of interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

1. Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased. ProPublica.
2. Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671–732.
3. Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. Proceedings of FAT, *149–159*.
4. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of FAT, 77–91.
5. Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163.
6. Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness. Proceedings of FAT, 643–669.
7. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. Proceedings of ITCS, 214–226.
8. Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. Proceedings of KDD, 259–268.
9. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems.
10. Kallus, N., Mao, X., & Zhou, A. (2020). Assessing algorithmic fairness with unobserved protected class using data combination. Management Science, 66(5), 1959–1981.
11. Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. Proceedings of ITCS.
12. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35.
13. Mitchell, S., Wu, S., Zaldivar, A., et al. (2019). Model cards for model reporting. Proceedings of FAT.
14. Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results. Proceedings of AAAI/ACM AIES.
15. Zliobaite, I. (2017). Measuring discrimination in algorithmic decision making. Data Mining and Knowledge Discovery, 31*(4)*, 1060–1089.