Available Online: 2 July 2025 DOI: 10.54254/3029-0880/2025.24261

Analysis and optimization of the applicability of hypothesis testing methods

Sirui Zhang

School of Mathematics and Computer Science, Hengshui University, Hengshui, China

hsxyjwc@hsnc.edu.cn

Abstract. With the continuous development of science and technology and other fields in today's world, statistical analysis and research have become indispensable research methods in people's daily lives. Among these, hypothesis testing plays an important role in fields such as biology, medicine, and economics, and has significant effects in most scenarios. However, the suitability and effectiveness of different hypothesis testing methods vary depending on the context, often leading to different outcomes and levels of accuracy. This paper focuses on discussing various hypothesis testing methods in statistics, such as t-test, chi-square test, z-test, F-test, etc. In addition, this study analyzes various cases in real life and optimize and improve some hypothesis testing methods. Based on a review of existing literature, the study explores how traditional hypothesis testing determines statistical significance—where the null hypothesis is rejected if the test statistic falls within the rejection region. To improve and enhance the determination of the significance level, handle uncertainty, and increase the sample size, this paper proposes alternative methods such as NHST test, Bayesian test, big data sequential test, and failure rate hypothesis test from the perspectives of medicine and kinesiology.

Keywords: hypothesis testing, applicability analysis, significance test, index evaluation method, multiple hypothesis testing

1. Introduction

Hypothesis testing is an important part of inferential statistics. It uses sample information to determine whether the assumptions about the population parameters or the population distribution are credible. It includes parametric tests and non-parametric tests. The most commonly reported statistical test results are based on the significance test of the null hypothesis [1]. The core principle underlying hypothesis testing lies in the concept of small probability—that is, assessing whether an observed result is likely to have occurred by chance in order to evaluate the validity of an assumption [2].

Hypothesis testing is widely applied across various fields. For instance, in the aviation industry, it serves as a critical step in the development and performance validation of aviation equipment, where product indicators must be rigorously tested. In the medical field, after performing descriptive statistical analysis on variables, formal hypothesis testing is essential to determine whether observed differences in treatment effects or group comparisons are statistically significant or merely due to random variation [3].

This paper focuses on evaluating the limitations and differences in the applicability of various hypothesis testing methods. It aims to compare the strengths and weaknesses of different test types and propose potential improvements. The findings contribute to the theoretical development of statistical hypothesis testing and partially address existing gaps in the current research landscape.

2. Hypothesis testing

2.1. Basic concepts

Hypothesis testing, also known as statistical hypothesis testing, is a statistical inference method used to determine whether differences between samples or between a sample and a population are due to sampling error or inherent discrepancies. The basic steps include proposing the null hypothesis (H0) and alternative hypothesis (H1), selecting a significance level, calculating the test statistic, and making decisions based on the p-value or critical value. Common testing methods include the t-test, z-test, chi-square test, and F-test.

Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

2.2. Theoretical foundation

Hypothesis testing is based on the principle of small probability, which evaluates the validity of a hypothesis by observing whether a low-probability event occurs. The process begins by proposing H0, defining H1 as the opposite of H0, and setting α (the threshold for small probability events). A statistical method is then applied to assess the likelihood of H0 being true. If this likelihood is sufficiently small, H0 is rejected; otherwise, it is retained [2].

2.3. Common methods and applicable conditions

t-test: Compares means between two groups; suitable for small samples with unknown population variance.

z-test: Used for large samples with known population variance.

Chi-square test: Tests independence or goodness-of-fit for categorical data.

F-test: Utilized in analysis of variance (ANOVA) to compare means across multiple groups.

2.4. Two types of errors in hypothesis testing

Hypothesis testing may lead to two errors:

Type I error (False Positive): Rejecting H0 when it is actually true, due to sampling randomness placing the sample in the rejection region. The probability of this error is denoted as α.

Type II error (False Negative): Failing to reject H0 when it is false, due to sampling randomness placing the sample in the acceptance region. The probability of this error is denoted as β.

Under fixed sampling conditions, reducing both errors simultaneously is impossible (decreasing one increases the other). The only way to minimize both is to increase the sample size, which highlights a key limitation of hypothesis testing [2].

2.5. Debates and challenges

Hypothesis testing remains a topic of debate despite deeper understanding of its principles. Key controversies include disputes over null hypothesis significance testing (NHST) and the misuse of p-values. A common misunderstanding lies in conflating the conditional probability p(data | H0) (probability of observing the data assuming H0 is true) with p(H0 | data)(probability of H0 being true given the data). Additionally, large sample sizes often yield extremely small p-values, leading to overinterpretation. To address these issues, researchers continually refine and improve hypothesis testing methodologies [4].

2.6. Optimization and alternative methods

Failure Rate Hypothesis Testing

Failure rate hypothesis testing, grounded in goodness-of-fit tests and parameter estimation, offers a robust alternative to traditional statistical hypothesis testing. It features simple parameters, mathematical tractability, broad applicability, and effective risk cost control.

Case Study: Ice Detector Failure Rate Evaluation

Using operational failure data from a similar aircraft model, we assessed the failure rate of ice detectors. Historical fault data for the aircraft's ice detectors were cleaned and processed, revealing the device's failure patterns (as illustrated in Table 1 below):

Table 1. The historical record of failures of the ice detector

Time	Monthly Flight Time (hours)	Frequency	Failure Rate (%)
January 2012	77.51648	3	38.70
February 2012	70.92096	1	-
March 2012	69.12112	2	28.93
April 2012	70.12192	2	28.52
May 2012	78.06912	2	25.62
June 2012	76.69216	4	52.16
July 2012	77.00360	1	-
August 2012	78.92336	3	38.01
September 2012	83.59384	1	-
October 2012	84.76240	2	23.60
November 2012	79.85496	2	25.05

Table 1. Continued

December 2012	78.43720	3	38.25
January 2013	78.64148	3	38.15
February 2013	72.04596	1	-
March 2013	70.24612	2	28.47
April 2013	71.24692	1	-
May 2013	79.19412	4	50.51
June 2013	77.81716	2	25.70
July 2013	78.12860	3	38.40
August 2013	80.04836	2	24.98
September 2013	84.71884	2	23.61
October 2013	85.88740	2	23.29
November 2013	80.97996	3	37.05
December 2013	79.56220	2	25.14
January 2014	78.40448	2	25.51
February 2014	71.80896	2	27.85
March 2014	70.00912	2	28.57
April 2014	71.00992	3	42.25
May 2014	78.95712	3	38.00
June 2014	77.58016	2	25.78
July 2014	77.89160	3	38.52
August 2014	79.81136	2	25.06
September 2014	84.48184	3	35.51
October 2014	85.65040	2	23.35
November 2014	80.74296	3	37.15
December 2014	79.32520	2	25.12

Note: "**-**" indicates that the data for that month does not meet the criteria, and the component failure rate for that month is excluded from calculation.

For the hypothesis testing on the failure rate of the ice detector, the null and alternative hypotheses are:

1) Null hypothesis (H0): $\lambda \leq 38.051 \times 10^{-6}$ / FH (i.e., the ice detector provided by the supplier meets the failure rate requirement).

2) Alternative (research) hypothesis (H1): $\lambda > 38.051 \times 10^{-6}$ / FH (i.e., the ice detector provided by the supplier fails to meet the failure rate requirement).

A significance level of $\alpha = 0.05$ is selected, resulting in a rejection region for the test statistic z of $z > 1.645$.

At the significance level $\alpha = 0.05$, the calculated test statistic z (-0.56) does not fall within the rejection region ($z > 1.645$). Therefore, we fail to reject H0 and maintain the status quo until compelling evidence emerges to refute H0, i.e., the ice detector's failure rate meets its design requirements during the aircraft development phase [5].

However, when reviewing monthly sample values in Table 2, seven observations exceed the design threshold, while 24 fall below it. Despite this, raw data distribution alone is insufficient to support definitive conclusions. This exemplifies the classic statistical inference challenge—drawing population-level conclusions from sample data.

Assessing performance metrics (e.g., failure rates) based on operational, flight test, or experimental fault data represents a classic scenario of inferring population characteristics from samples. However, directly extrapolating sample statistics to the population inevitably introduces errors. In civil aircraft performance evaluation, the null hypothesis is typically set as compliance with requirements. For "smaller-is-better" metrics (e.g., failure rates), an upper-tailed test is preferred, while a lower-tailed test is used for "larger-is-better" metrics.

3. Practical applicability analysis

3.1. Hypothesis testing in sports strategy evaluation

Hypothesis testing can analyze tactical strategies in sports. For example, in a badminton match analysis:

Collect match footage of players A and B.

Define variables such as rally length and diagonal shot frequency.

Propose hypotheses:

H1: Player A uses fewer diagonal shots than the opponent but scores more points with diagonals.

H2: Player B excels in winning points through extended rallies.

Using hypothesis testing metrics, the analysis reveals that while Player A's diagonal shots increase the opponent's movement and fatigue, they also demand higher physical stamina from Player A. Tactical elements should therefore be analyzed within integrated models that support broader strategic frameworks rather than as isolated variables [6].

3.2. Hypothesis testing in medical research

Medical statistics, a critical interdisciplinary tool, employs hypothesis testing to:

Evaluate new drug efficacy.

Assess treatment effectiveness.

Analyze disease risk factors.

Validate diagnostic test accuracy.

For instance, in a study testing the hypothesis "Drug A significantly reduces blood pressure," a randomized controlled trial divides patients into two groups: one receiving Drug A and the other a placebo. Results show Drug A's group has a significantly lower mean blood pressure ($p < 0.05$), confirming its effectiveness [3].

4. Conclusion

This study focuses on the applicability analysis and optimization of hypothesis testing methods, systematically evaluating and refining various approaches. The conclusions align with traditional hypothesis testing: rejecting H0 if the test statistic falls in the rejection region; otherwise, failing to reject H0. However, several limitations remain. In particular, the validity of conclusions in medical research is highly contingent on the accuracy of the null hypothesis. Mis-specification of H_0may lead to erroneous interpretations.To address these challenges, future statistical analyses should prioritize data accuracy and hypothesis specificity to systematically resolve related issues.

References

[1] Wen, Z. L., Xie, J. Y., Fang, J., & Wang, Y. F. (2022). Methodological research on hypothesis testing and related issues in China over the past two decades. *Advances in Psychological Science, 30*(8), 1667–1681.

[2] Gong, J. J. (2023). Research and practical exploration of IFHCI quality audit methods based on statistical hypothesis testing theory. *Value Engineering, 42*(27), 33–35.

[3] Zeng, H. X., Yang, Z. Y., Liu, D. H., Wang, R. H., Chen, H. S., Zhang, H. W., . . . & Cao, G. W. (2023). Application of common statistical analysis methods in medical research. *Shanghai Journal of Preventive Medicine, 35*(8), 831–839. https://doi.org/10.19428/j.cnki.sjpm.2023.22771

[4] Wen, Z. L., Xie, J. Y., Fang, J., & Wang, Y. F. (2022). Methodological Research on Hypothesis Testing and Related Issues in China over the Past Two Decades. *Advances in Psychological Science, 30*(08), 1667–1681.

[5] Zheng, Y. L. (2024). Failure rate evaluation method based on fault data and hypothesis testing. *Civil Aircraft Design and Research, 4*, 71–75. https://doi.org/10.19416/j.cnki.1674-9804.2024.04.012

[6] Xu, X. Q., Guo, P. C., Feng, G. S., Wang, H. W., Liu, Z. M., & Han, W. (2023). Application of hypothesis testing methods in technical and tactical analysis of elite badminton matches. In *Proceedings of the 13th National Sports Science Congress: Special Report (Sports Statistics Section)* (pp. 87–89). Competitive Sports College of Shandong Sport University; Qixian Second Senior High School Sports Group; School of Physical Education, Shandong Sport University.