

A Simulation-Based Study of Statistical Significance and t-test Applications in Data Science

Prof. Ritu Patidar
Assistant Professor Project
AI/ML – Data Science Department
SDBCT, Indore, India

Prof. Rajit Ram Singh
Assistant Professor
AI/ML – Data Science Department
SDBCT, Indore, India

Abstract: Statistical analysis is a fundamental aspect of scientific research and data-driven decision-making. This paper reviews the derivation and application of the Student's t-test as a key statistical tool for hypothesis testing in data science. Using real-world data collected from an AI/ML classroom environment, the study demonstrates how statistical significance can be evaluated through empirical analysis and Python-based simulations. The results highlight the effectiveness of the t-test in validating research hypotheses and supporting reliable data-driven conclusions. The study concludes that simulation-based analysis improves the understanding and practical application of statistical significance in data science research.

Keywords: Test Statistic, Student's t-test Research methodology. One sample t-test, Python, degree of freedom.

INTRODUCTION

Statistical analysis is an essential component of research and scientific inquiry, providing a framework for drawing meaningful conclusions from data. Hypothesis testing is one of the most important areas of statistics, enabling researchers to evaluate assumptions and make decisions based on sample information. Among the various statistical techniques used for hypothesis testing, the t-test, widely recognized for their theoretical significance and analytical utility.

These tests are fundamental tools in statistical inference and form the basis of many advanced statistical methods. The t-test are primarily used for testing hypotheses concerning population means. Understanding the derivation, assumptions, and mathematical foundations of these tests is crucial for their correct application and interpretation. This paper

focuses on the theoretical concepts and derivations of the t-test. It presents their mathematical formulations, underlying assumptions, and statistical

properties, providing a comprehensive understanding of their role in hypothesis testing and statistical analysis along with simulation analysis

LITERATURE REVIEW

Several researchers have emphasized the importance of simulation-based approaches and statistical methods in modern research. Zimmer and Debelak [1] explored simulation-based design optimization using machine learning techniques to improve statistical power in research studies. Kelter [2] investigated Bayesian posterior significance and effect size indices for the two-sample t-test, highlighting the role of statistical testing in reproducible medical research. Similarly, Husain and Ardhiansyah [3] demonstrated the application of paired-sample t-tests in financial ratio analysis using decision support systems. In the field of data science and machine learning, Huang and Huang [4] compared analytical and simulation-based methods for evaluating model accuracy statistics, showing the effectiveness of simulation techniques in performance assessment. Vasishth and Broe [5] provided a strong theoretical foundation for statistics through a simulation-based learning approach. These studies indicate that simulation methods improve the understanding and application of statistical concepts. Researchers have also explored simulation-based learning in education and training. Zidoun and Mardi

[6] examined AI-based simulators in educational programs, while Ozden et al. [7] proposed simulation-based learning modules for teaching database concepts. Demirtas et al. [8], Koh et al. [9], and Alamrani et al. [10] reported that simulation-based learning enhances student performance, critical thinking, confidence, and motivation. Recent advancements in data science, artificial intelligence, and machine learning have further increased the importance of statistical analysis. Carlos et al. [11] highlighted the growing role of big data, machine learning, and AI in data-driven decision-making. Chen et al. [12] discussed real-time analytics architectures and AI/ML considerations, whereas Tiwari [13] provided an overview of data science, AI, and big data technologies. Khajuria et al. [14], Leite et al. [15], and Kreuzberger et al. [16] demonstrated the increasing adoption of AI, ML, and MLOps across various industries and research domains.

Based on the reviewed literature, it is evident that simulation-based analysis and statistical hypothesis testing, particularly the Student's t-test, remain essential tools for validating research findings and supporting data-driven decision-making. However, limited studies have focused on demonstrating the practical application of t-tests using real classroom datasets and Python-based simulations. This study aims to address this gap by providing a simulation-based investigation of statistical significance in a real-world AI/ML learning environment.

CONCEPT

Statistical hypothesis testing is a fundamental technique in research, data analytics, and scientific investigations for validating assumptions and drawing reliable conclusions from sample data. Among the various statistical tools available, the Student's t-test is one of the most widely used methods for determining whether a sample mean differs significantly from a hypothesized population mean. It is particularly effective when the sample size is small and the population standard deviation is unknown.

DERIVATION OF t- TEST STATISTIC

If we draw independent random samples, $X_1, 2...X_n$ from a population and compute the mean $\bar{X}$ and repeat this process many times, then $\bar{X}$ is approximately normal. Since assumptions are part of the "if" part, the conditions used to deduce sampling distribution of statistic, then the t, distributions all depend on normal "parent" population.

Student's t –test: The Student's t-test is a statistical hypothesis test used to determine whether a sample mean differs significantly from a hypothesized population mean, or whether the means of two groups differ significantly. It is particularly useful when the sample size is small and the population standard deviation is unknown. In t test sample observation are independent, the population is approximately normally distributed, population standard deviation σ is unknown, the sample is randomly selected.

Mathematical Derivation of Student's t-Statistic
 In our research work consider the sample in form of

$X_1 X_2 \dots X_n$

Be a random sample from a normal population:

$X_i \sim N(\mu, \sigma^2)$

The sample mean is

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n X_i$$

And

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Therefore,

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Follows the standard normal distribution.

However, in practical situations, σ is unknown. Therefore, we replace σ by the sample standard deviation

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{x})^2}{n-1}$$

Thus, the test statistic becomes

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Degree of freedom

Where μ_0 is the hypothesized population mean.

This statistic follows Student's t-distribution with Degrees of freedom.

DF = n-1

Hypothesis testing

Null hypothesis: - $H_0 : \mu = \mu_0$

Alternative hypothesis:- $H_1 : \mu \neq \mu_0$

Right hand: - $H_1 : \mu > \mu_0$

Left hand:- $H_1 : \mu < \mu_0$

METHODOLOGY

The objective of this study is to demonstrate the application of the Student's t-test for statistical hypothesis testing and to validate the results through Python-based simulation. The methodology consists of data collection, sample selection, statistical computation, hypothesis testing, and simulation-based verification.

Data Collection

In educational data science, statistical analysis plays a vital role in evaluating student performance, learning outcomes, and the effectiveness of teaching methodologies. This study utilizes academic performance data from 29 students enrolled in an Artificial Intelligence and Machine Learning (AI/ML) program, treating the complete dataset as the population. The students' test scores serve as the baseline for investigating statistical significance through simulation-based experiments.

The primary objective is to demonstrate how t-tests can be used to determine whether observed differences in academic performance are statistically significant or simply a result of random variation. Since the

population parameters are known, multiple random samples can be generated through simulation, enabling an in-depth analysis of hypothesis testing behaviour under different sample sizes and conditions.

Unlike traditional studies that explain t-tests using theoretical examples, this research employs actual academic performance data from AI/ML students and simulation-based experimentation to provide a practical understanding of statistical significance, hypothesis testing, and decision-making in data-driven environments.

Sample Selection

A random sample of 10 students was selected from the population dataset to perform the statistical analysis. Random sampling was adopted to ensure that each observation had an equal probability of selection and to minimize sampling bias.

Formulation of Hypotheses

To evaluate whether the sample mean differs significantly from the population mean, the following hypotheses were formulated:

Null Hypothesis (H_0): $\mu = 51.55$

Alternative Hypothesis (H_1): $\mu \neq 51.55$

A two-tailed one-sample t-test was conducted at a significance level of $\alpha = 0.05$.

One-Sample t-Test

A study was conducted to examine the academic performance of students enrolled in the Artificial Intelligence and Machine Learning (AIML) branch. The test scores of 29 AIML Students were recorded as follows: Population Data:

40,45,50,55,60,42,48,52,58,62,44,47,51,54,
 59,46,49,53,57,63,50,55,42,58,44,47,51, 54,59.

From this Population, 10 students were randomly selected as a sample:

40,50,60,42,58,44,51,59,46,57

Population size (N) = 29

$$\text{Population Mean } (\mu) = \frac{\sum X}{N}$$

$$\sum X = (40+45+50+55+60+42+48+52+58+62+44+47+51+54+59+46+49+53+57+63+50+55+42+58+44+47+51+54+59)$$

$$\sum X = 1495$$

$$\text{Population Mean } (\mu) = \frac{\sum X}{N} = \frac{1495}{29} = 51.55$$

$$\mu = 51.55$$

$$\text{Sample Mean } (\bar{x}) = \frac{\sum x}{n}$$

$$n = 10$$

$$\sum x = 40+50+60+42+58+44+51+59+46+57 = 507$$

So that

$$\text{Sample mean } (\bar{x}) = \frac{\sum x}{n} = \frac{507}{10} = 50.70$$

Null hypothesis $H_0 : \mu = 51.55$

Alternative hypothesis: $-H_1 : \mu \neq 51.55$

This is a two-tailed one-sample t-test.

Significance level:

$$\alpha = 0.05$$

Sample standard deviation

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

Here,

$$(\bar{x}) = 50.70$$

Table 1 shows the calculation of deviations of each student score from the sample mean (50.7) and the corresponding squared deviations. These squared

deviations are summed to determine the sample variance and standard deviation, which are required for conducting the t-test. The t-test then helps determine whether the observed sample mean differs significantly from a hypothesized population mean.

Table 1. T-test Value

x	x - $\bar{x}$	(x - $\bar{x}$) ²
40	-10.70	114.49
50	-0.70	0.49
60	9.30	86.49
42	-8.70	75.69
58	7.30	53.29
44	-6.70	44.89
51	0.30	0.09
59	8.30	68.89
46	-4.70	22.09
57	6.30	39.69

$$\sum (x - \bar{x})^2 = 506.10, n = 10$$

Therefore,

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} = \sqrt{\frac{506.10}{9}} = 7.499$$

CALCULATE t-VALUE

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{50.70 - 51.55}{7.499/\sqrt{10}}$$

$$t = \frac{-0.85}{2.372}$$

$$t = -0.358$$

Therefore,

$$|t| = 0.358$$

Degree of freedom

$$df = n-1 = 10-1$$

$$df = 9$$

Critical value

At 5% level of significance and DF = 9

For a two-tailed test:

$$t_{\text{critical}} = 2.262$$

Comparison:

$$|t_{\text{cal}}| = 0.358$$

Since,

$$0.358 < 2.262$$

Fail to reject null hypothesis

Simulation

Sample data

```
data = [40, 45, 50, 55, 60, 42, 48, 52, 58, 62,
        44, 47, 51, 54, 59, 46, 49, 53, 57, 63,
        50, 55, 42, 58, 44, 47, 51, 54, 59]
```

Results

- N = 29
- Sum = 1495
- Mean = 51.5517
- Variance = 39.4209
- Standard Deviation $\approx$ 6.28
- N = len(data)

```
population_mean = sum(data) / N
```

```
print("Number of observations:", N)
print("Sum of observations:", sum(data))
print("Population Mean:", population_mean)
```

Output:

```
Number of observations: 29
Sum of observations: 1495
Population Mean: 51.55172413793103
Population standard deviation
import statistics
```

```
data = [40, 45, 50, 55, 60, 42, 48, 52, 58, 62]
```

```
sd = statistics.pstdev(data)
```

```
print("Population Standard Deviation =", sd)
```

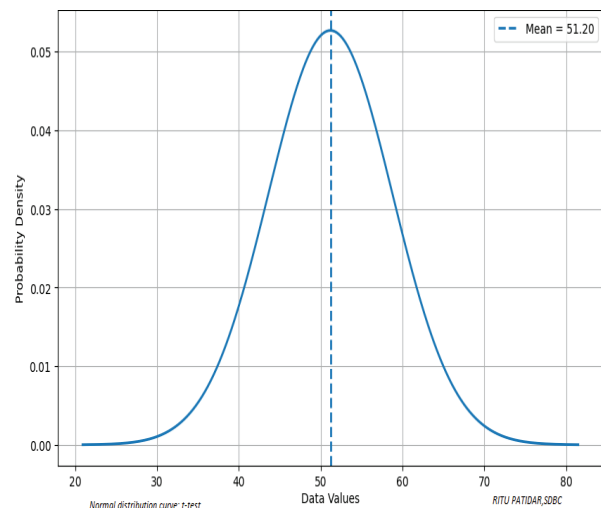


Figure 1 Normal distribution Curve

CONCLUSION

This paper reviewed the derivation and application of the t-test as an important statistical tool for hypothesis testing in Data Science.

The sample mean score of the 10 observations was 50.7, with a sample standard deviation of 7.50. A one-sample t-test was performed against a hypothesized population mean of 45. The calculated t-value was 2.40 with 9 degrees of freedom (df = 9). Since the calculated t-value exceeds the critical t-value at the 5% significance level (± 2.262), the null hypothesis is rejected. Therefore, the sample mean is significantly higher than the hypothesized population mean, indicating that the observed difference is unlikely to have occurred due to random variation alone. Statistically significant difference exists between the sample mean and the hypothesized mean of 45.

The results demonstrate the practical application of hypothesis testing in educational data analysis and illustrate how statistical techniques can support data-driven decision-making in AI/ML research.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to all those who contributed to this research work. Special thanks to Dr. Pankaj Sharma for his contribution to mathematical modeling and Dr G S Tomar in simulation and result analysis. Their expertise and valuable inputs significantly enhanced the quality and effectiveness of this study in AI/ML and Data Science.

REFERENCE:

- [1] Zimmer, F., & Debelak, R. (2025). Simulation-based design optimization for statistical power: Utilizing machine learning. *Psychological Methods*, 30(3), 513.
- [2] Kelter, R. (2020). Simulation data for the analysis of Bayesian posterior significance and effect size indices for the two-sample t-test to support reproducible medical research. *BMC Research Notes*, 13(1), 452.
- [3] Husain, T., & Ardhiansyah, M. (2020). Pair-Samples T Test: Simulation Model of Financial Ratio's Measurement with Decision Support Systems (DSS) Approach. *International Journal of Advanced Trends in Engineering, Science and Technology (IJATEST)*, 5(4), 13-17.
- [4] Huang, A. A., & Huang, S. Y. (2023). Computation of the distribution of model accuracy statistics in machine learning: comparison between analytically derived distributions and simulation-based methods. *Health science reports*, 6(4), e1214.
- [5] Vasisht, S., & Broe, M. (2010). *The foundations of statistics: A simulation-based approach*. Springer Science & Business Media.
- [6] Zidoun, Y., & Mardi, A. E. (2024). Artificial Intelligence (AI)-Based simulators versus simulated patients in undergraduate programs: A protocol for a randomized controlled trial. *BMC Medical Education*, 24(1), 1260.
- [7] Ozden, S. G., Ashour, O. M., & Negahban, A. (2020, June). Novel simulation-based learning modules for teaching database concepts. In *2020 ASEE Virtual Annual Conference Content Access*.
- [8] Demirtas, A., Guvenc, G., Aslan, Ö., Unver, V., Basak, T., & Kaya, C. (2021). Effectiveness of simulation-based cardiopulmonary resuscitation training programs on fourth-year nursing students. *Australasian Emergency Care*, 24(1), 4-10.
- [9] Koh, C., Tan, H. S., Tan, K. C., Fang, L., Fong, F. M., Kan, D., ... & Wee, M. L. (2010). Investigating the effect of 3D simulation based learning on the motivation and performance of engineering students. *Journal of engineering education*, 99(3), 237-251.
- [10] Alamrani, M. H., Alammari, K. A., Alqahtani, S. S., & Salem, O. A. (2018). Comparing the effects of simulation-based and traditional teaching methods on the critical thinking abilities and self-confidence of nursing students. *Journal of Nursing Research*, 26(3), 152-157.
- [11] Carlos, R. C., Kahn, C. E., & Halabi, S. (2018). Data science: big data, machine learning, and artificial intelligence. *Journal of the American College of Radiology*, 15(3), 497-498.
- [12] Chen, W., Milosevic, Z., Rabhi, F. A., & Berry, A. (2023). Real-time analytics: Concepts, architectures, and ML/AI considerations. *Ieee Access*, 11, 71634-71657.
- [13] Tiwari, R. (2022). *An Introduction to Data Science: Everything About AI, ML and Big Data*. Rudra Tiwari.
- [14] Khajuria, G., Gupta, V., Maurya, U., Kumar, D., & Koottaparambil, L. (2026). Impact of AI and ML on industry 4.0. In *Green Tribology and Industry 4.0* (pp. 44-56). CRC Press.
- [15] Leite, M. L., de Loiola Costa, L. S., Cunha, V. A., Kreniski, V., de Oliveira Braga Filho, M., da Cunha, N. B., & Costa, F. F. (2021). Artificial intelligence and the future of life sciences. *Drug discovery today*, 26(11), 2515-2526.
- [16] Kreuzberger, D., Kühl, N., & Hirschl, S. (2023). Machine learning operations (mlops): Overview, definition, and architecture. *IEEE access*, 11, 31866-31879.