Journal homepage: https://itce.com.ua/en

Information Technologies and Computer Engineering **Vol. 21, No. 3. 2024**

Article's History: Received: 03.08.2024 Revised: 19.10.2024 Accepted: 26.12.2024

UDC 519.23:004.021 10.63341/vitce/3.2024.119

Improved A/B testing acceleration methods for parametric hypothesis testing: T-test comparison with CUPED, CUPED++ and Bayesian Estimator

Artur Markov*

Postgraduate Student
Taras Shevchenko National University of Kyiv
01033, 60 Volodymyrska Str., Kyiv, Ukraine
https://orcid.org/0009-0000-5222-4397

Abstract. The study aimed to compare statistical analysis methods to improve the testing of alternatives. The study evaluated four main methods: the classic T-test, the conventional and advanced method of Controlled Experiments Using Pre-Experimental Data (CUPED), and the Bayesian Estimator. The main results included a demonstration of the A/B testing process, and the described statistical analysis methods included detailed characteristics and examples of use. The simulations and practical application revealed that the T-test provides high accuracy with small samples, but its effectiveness decreases with increasing sample size due to high resource requirements. The calculator for this method demonstrated effectiveness in simple tasks but had limitations with large data. The conventional CUPED method has shown increased accuracy due to variation correction, but its effectiveness decreases when working with large and complex data sets. The written program for this method has shown to be effective in cases where the previous data is well represented, but its capabilities are limited when processing large data sets. The improved version provided a significant improvement in both accuracy and processing speed, especially for large datasets, thanks to advanced modelling and optimisation. The code results confirmed that this method is highly efficient for complex experiments, particularly when processing large amounts of data. Moreover, the Bayesian Estimator demonstrated high accuracy due to the integration of prior knowledge but required more computational resources and time. The platform used for this method demonstrated the ability to account for uncertainty yet required complex model settings. The results highlighted the importance of selecting the appropriate statistical analysis method depending on the scale and complexity of the data to ensure optimal accuracy and efficiency of testing

Keywords: statistical analysis; correction of variations; effectiveness of approaches; data processing; modelling of experiments

Introduction

With the rapid development of information technology and the increase in data volumes, the effectiveness of statistical analysis methods has become highly relevant. One of the basics of statistical analysis is A/B testing, which is used to compare two or more options and determine the most effective one. The basics of such testing include comparing the mean values in two groups using parametric hypotheses to assess the statistical significance of differences between the groups. Traditionally, such comparisons are made using the T-test, which is a method for analysing average values. However, to improve accuracy and speed, advanced methods have been developed, such as Controlled Experiments Using Pre-Experimental Data (CUPED) and its extended version CUPED++, as well as the Bayesian Estimator method.

The T-test method, although a basic tool for comparing values, has limitations when analysing large and complex data sets due to high resource requirements and inaccuracy in cases of variation. Although CUPED can be used to correct for variations by using previous data, its effectiveness decreases when processing large amounts of data. The extended version of CUPED++ offers certain improvements but still requires a detailed comparison with other modern methods, such as the Bayesian Estimator, which provides high accuracy by integrating prior knowledge but is

Suggested Citation:
Markov, A. (2024). Improved A/B testing acceleration methods for parametric hypothesis testing: T-test comparison with CUPED, CUPED++ and Bayesian Estimator. *Information Technologies and Computer Engineering*, 21(3), 119-131. doi: 10.63341/vitce/3.2024.119

*Corresponding author

resource-intensive. Existing studies do not sufficiently address these aspects, which determines the need for a deeper analysis and comparison of new and improved methods in the context of their practical application.

For instance, V. Kalchenko (2018) analysed common methods of testing for penetration into computer systems and proposed a classification of these methods, noting their advantages and disadvantages. O.V. Shportko & M.M. Mushyn (2023) substantiated the effectiveness of the method of gradual formation of a set of objective function values for combinatorial optimisation problems, demonstrating that this method significantly reduces the execution time of programs compared to other approaches, such as search methods with returns. In addition, V. Khambir (2024) proved that the automation of testing processes, including functionality, interface, performance, and security testing, facilitates and increases the efficiency of application testing, despite the problems with the initial setup and complexity of testing complex scenarios.

Furthermore, K. Kachiashvili (2018) described the intensive development of statistical hypothesis testing methods and introduced a new approach known as Constrained Bayesian Methods to solve complex problems in various fields, including a detailed description of existing methods, new approaches and special software for their applications. X. Gu *et al.* (2023) presented a Bayesian method for testing information hypotheses in confirmatory factor analysis models, in particular, by estimating factor loadings using Bayesian factors and posterior probabilities of models, which can be used to solve various issues, including the assessment of equality of loadings and indicator validity. Additionally, Ramesh *et al.* (2023) reviewed the process of formulating and testing hypotheses, emphasising the importance of theoretical grounding and a systematic approach to hypothesis testing, which ensures objectivity in the search for knowledge. A. Deng *et al.* (2021) developed a new unbiased reduced variance estimator of intervention effects that uses the idea of improving the efficiency of CUPED for cases where the trigger status is observed only in a certain group, which allows achieving accuracy comparable to full status observation.

In turn, J.J. Wang (2022) developed an approximate Bayesian estimator for the parameters of a regression model with random coefficients, which outperforms traditional testing methods in terms of estimation accuracy, simplifying the calculation and interpretation of results. S. Woo (2023) presented parametric accelerated durability testing as a systematic method for assessing the reliability of mechanical system structures, which allows the detection of structural defects and reduces the failure rate due to material fatigue. In addition, Y.W. Cho *et al.* (2024) presented the results of a series of Monte Carlo simulations to evaluate methods for fitting multilevel models of latent differential structural equations, finding that the Bayesian approach outperforms frequency-based methods, especially in the context of handling variable covariates and coupling parameters.

Although these studies addressed various statistical methods, there are still gaps in their adaptation to big data and complex experiments. This study describes and compares the effectiveness of such methods in improving A/B testing, with a focus on the accuracy and speed of data analysis. The tasks included a theoretical study of the basics of A/B testing and advanced methods, conducting simulations to assess their practical applicability, and analysing the effectiveness of different approaches.

Materials and Methods

The study was initiated by a comprehensive analysis of the theoretical context of A/B testing and improved analysis methods, including the basics and value of this testing for comparing alternatives. Parametric hypotheses, their role in testing, and their impact on the accuracy and speed of results were studied. Traditional and advanced methods, such as T-test, CUPED, CUPED++, and Bayesian Estimator, their algorithms and applications, as well as the importance of advanced methods for improving testing efficiency were highlighted.

A detailed description and characterisation of these methods of statistical analysis were provided in the study. For each method, the principles of operation, implementation of algorithms and examples of application are described. The classical T-test was considered a basic method for comparing mean values, CUPED and CUPED++ as methods for correcting variations to improve the accuracy of results, and the Bayesian Estimator as a method for estimating probabilities with the integration of prior knowledge. The practical application of the methods included the creation and testing of code and formulas for each method. The general formula for the T-test method (1):

$$T = \frac{M_1 - M_2}{SE}, \quad (1)$$

where: M_1, M_2 – two mean values; SE – overall standard error for the two samples.

The general formula for the CUPED method is (2):

$$\hat{Y}_{cv} = \bar{Y} - \theta\bar{X} + \theta E(X), \quad (2)$$

where: X – value before the experiment; Y – value of the experiment; θ – a constant derived from the data that determines how much the results of the experiment should be adjusted based on the previous data; $\overline{Y}$ – average value of the experiment results Y; $\overline{X}$ – average value of the previous data X; $E(X)$ – mathematical expectation of the values X which is used for correction.

The general formula for the Bayesian Estimator method (3):

$$P(A|B) = \frac{P(B|A)*P(A)}{P(B)}, \quad (3)$$

where: $P(A|B)$ – the probability that a certain state or property (A) exists, given another property or condition (B); $P(B|A)$ – the probability that a certain property or condition (B) exists, given that another state or property

(*A*) exists; *P*(*A*) – the probability that (*A*) exists without taking into account additional conditions or properties; *P*(*B*) – the probability of having (*B*) without taking into account additional conditions or properties.

The programs for the implementation of CUPED and CUPED++ were written in Python using the pandas and stats models libraries and tested on the Replit platform. In turn, an online calculator was used to demonstrate the T-test method, and the Bayesian Estimator was based on the platform "Bayesian Estimation Supersedes the T-test – online", which shows how much better this method is than the T-test.The comparison of the effectiveness of the methods was based on an assessment of the accuracy, speed and overall efficiency of each method. The results were analysed to determine the advantages and limitations of each method, which was used to provide recommendations for practical application depending on the specifics of the tasks and the size of the data sets.

Results

Theoretical context of A/B testing and advanced analysis methods

A/B testing is a standard method for comparing the effectiveness of two or more alternatives in a statistical analysis. This approach is used to determine which of the options (A or B) is more effective in achieving a particular goal or indicator. A/B testing is based on the idea of random experiments, where respondents are randomly assigned to different groups, which minimises systematic errors and reveals the net effect of changes.

This testing usually involves two stages: dividing users into control (A) and experimental (B) groups and evaluating the results to compare effectiveness (Fig. 1). The main statistical tool for analysing the results of A/B testing is statistical hypothesis testing, which can be used to determine whether the difference between the groups is statistically significant.

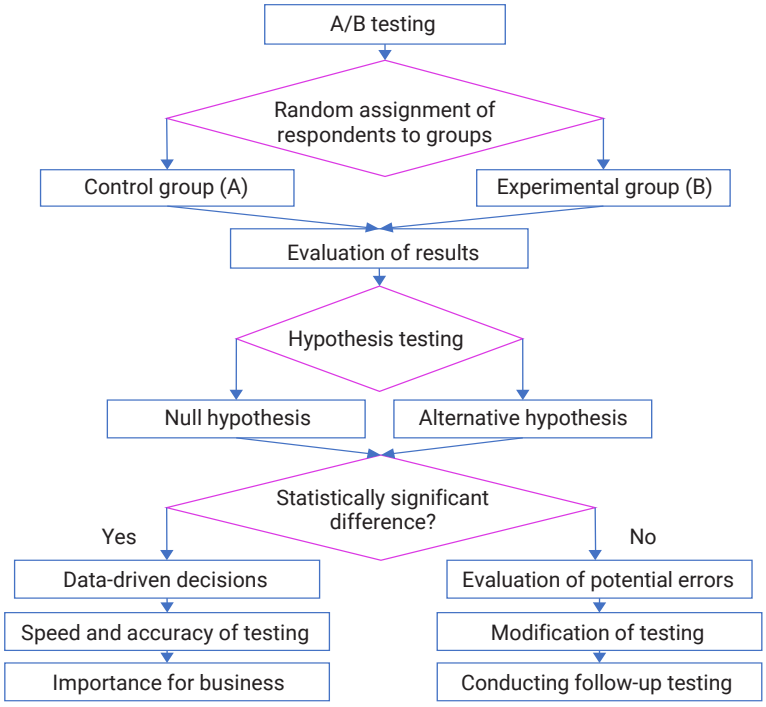

Figure 1. Flowchart of the A/B testing process

Source: compiled by the author

In addition, A/B testing can be used to objectively compare alternatives and make informed decisions based on empirical data. This is important for businesses where even small improvements can have a significant impact on the performance of campaigns or products. The success of A/B testing depends on the ability to evaluate results quickly and accurately. Increasing the speed and accuracy of testing improves operational decision-making based on up-to-date data and minimises the risk of false conclusions. This is especially relevant in a rapidly changing market and highly competitive environment. Also important are parametric hypotheses, which are mathematical models that formulate expectations about the distribution of data in the populations being compared. The null hypothesis (H0) typically states that there is no significant difference between the groups, while the alternative hypothesis (H1) states that there is a difference. In A/B testing, the correct formulation and testing of these hypotheses are critical to obtaining reliable results. The correct formulation and testing of parametric hypotheses affect the accuracy of testing. Incorrect assumptions can lead to erroneous conclusions and, as a result, to wrong decisions. In addition, methods that provide more accurate and faster results can significantly reduce the time required to collect and analyse data. Traditional

methods, such as T-test, provide a basic level of accuracy for A/B testing, but they have limitations in speed and accuracy, especially in the face of large amounts of data and frequent testing. Modern methods, such as CUPED, CUPED++ and Bayesian Estimator, provide improvements in these aspects with their advanced algorithms and techniques.

Improved analysis methods can reduce data variation, improve the accuracy of estimates and reduce the time required to reach statistical significance. For example, the CUPED method uses previous data to correct for variation and improve accuracy, while the Bayesian Estimator provides the flexibility to model and adapt to new data. These improvements contribute to more efficient testing and decision-making based on more reliable results.

Description and characteristics of statistical analysis methods

Statistical methods of analysis play a key role in A/B testing, providing effective comparisons of different alternatives and evaluation of their effectiveness. For this purpose, various methods are used, each of which has its characteristics and approaches to data processing. The standard T-test is the main statistical test for comparing the means of two independent groups. It determines whether there are significant differences between the means of the groups being compared. The basic principle of the t-test is to test H0, which states that the means of two groups do not differ from each other, against H1, which suggests that there is a statistically significant difference between the groups. The T-test procedure involves several key steps (Fig. 2). First, data are collected from two independent groups to be compared. Next, the means and standard deviations for each group are calculated. Based on these figures, the T statistic is calculated, which reflects the amount of difference between the means concerning the variation in the data. The test statistic T is compared with the critical value from the T-distribution, which assesses the probability of observing such a difference under the null hypothesis.

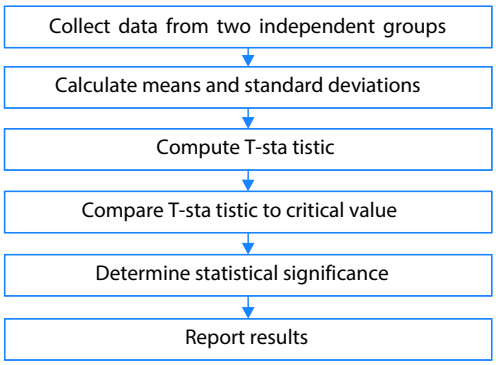

Figure 2. T-test scheme of operation
Source: compiled by the author

An example is the T-test, which examines the effectiveness of a new advertising message (group B) compared to an old advertising message (group A). If the results show that the average values of the performance indicators in groups A and B differ significantly, and this deviation is statistically significant, then it can be concluded that the new advert is effective.

The CUPED method, in turn, is a modern statistical approach that improves the accuracy of A/B testing by using previous data to correct for variations in results. It uses regression analysis to incorporate data collected before the experiment into the estimation of the effect of changes. This reduces the variation in test results and increases the accuracy of estimates. CUPED is based on two main stages (Fig. 3). The first stage involves collecting preliminary data, which may include information about user behaviour before the experiment or data about various user characteristics. In the second stage, this data is used to build a regression model that corrects for variations in the main results of A/B testing. Regression analysis can assess the impact of these variations on the final results, reducing noise and improving the accuracy of the estimates.

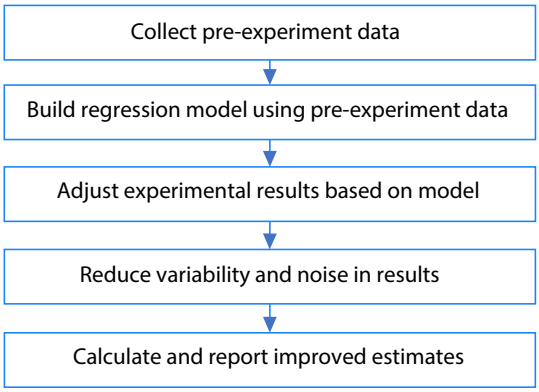

Figure 3. CUPED scheme of operation
Source: compiled by the author

For instance, if A/B testing is conducted to evaluate the effectiveness of a new feature in an application, CUPED can use data on user behaviour before the new feature is implemented to adjust the results. This provides a more accurate assessment of the impact of the new feature on user behaviour, as noise from various factors that affect the results will be reduced. Moreover, the CUPED++ method is an advanced version of CUPED designed to increase the accuracy of A/B testing by reducing variation and improving the correction of results. The main goal of CUPED++ is to further improve the process of variation correction by introducing new algorithms and advanced modelling methods. CUPED++ is based on CUPED but introduces several new optimisation approaches (Fig. 4). The main steps of CUPED++ include collecting preliminary data, building an extended regression model, and adapting and optimising the model for specific experimental conditions. CUPED++ uses more sophisticated regression algorithms that address additional factors and relationships between variables. This allows for even better correction of variations and reduces the impact of random noise on test results.

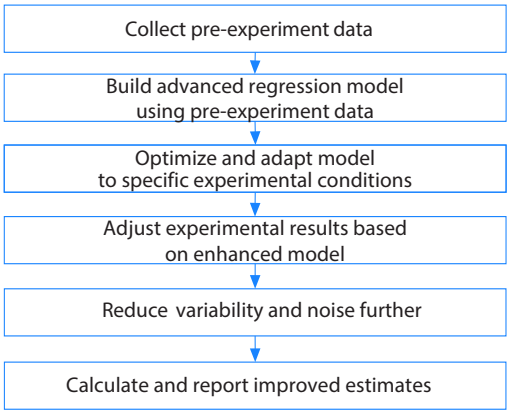

Figure 4. CUPED++ scheme of operation
Source: compiled by the author

For instance, when testing a new feature in a web application, CUPED++ can include not only data on user behaviour before the new feature was introduced but also data on seasonal trends or other factors that may affect the results. This provides even greater accuracy in measuring the effect of the new feature, as CUPED++ takes these additional variables into account when adjusting the results.

It is also worth considering the Bayesian Estimator, which is a Bayesian technique for estimating the difference between two means, which provides a probability distribution of this difference. Bayes' theorem describes how the probability of a certain hypothesis changes based on new information. The Bayesian Estimator is based on the concept of conditional probability, which integrates previous knowledge (a priori probabilities) and new observations to update probabilities (a posteriori probabilities). Bayesian analysis methods provide high flexibility in modelling and adapting to new data, which makes them useful in the face of uncertainty (Fig. 5). They can be used to estimate probabilities and draw reasonable conclusions using complex information from various sources. Bayesian methods are used for modelling and forecasting in the face of incomplete or inaccurate data, as well as for adapting to changes in data.

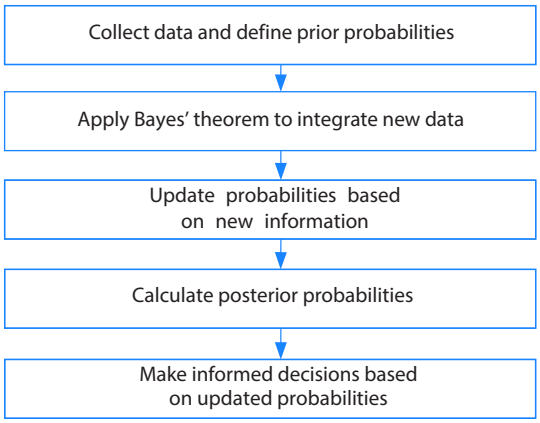

Figure 5. The Bayesian Estimator workflow
Source: compiled by the author

For instance, in A/B testing, a Bayesian Estimator can be used to determine the probability that option B is better than option A. If there are certain a priori assumptions about the effectiveness of options, the Bayesian approach adapts these assumptions based on the actual results of the experiment, providing more accurate and reasonable conclusions. Consequently, there are many different methods for evaluating and comparing alternatives, each with its characteristics and advantages. Existing approaches allow for precise data analysis, variation correction, adaptation to new conditions, and integration of prior knowledge with new data. The use of different methods makes it possible to choose the most effective tools for specific tasks, which is critical for achieving accurate and reliable results in testing and analysis.

Simulations and practical application of statistical methods

In general, the T-test can be used to compare mean values between two groups. There are several types of T-tests for two samples, such as unpaired, Welch's, and paired T-tests. In addition, there are many platforms and calculators that can calculate the necessary parameters to select the most appropriate test for analysis (Fig. 6).

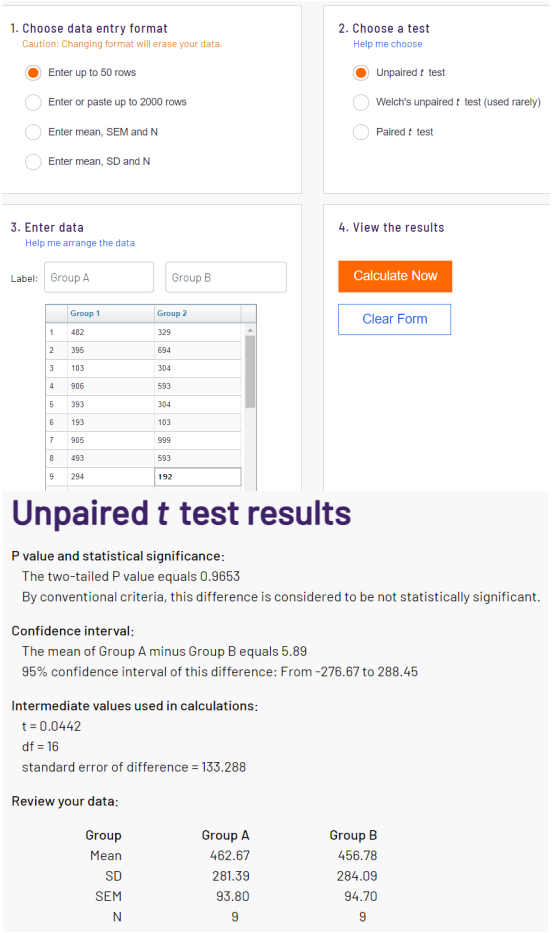

Figure 6. An example of using a calculator for a T-test
Source: compiled by the author

This test focuses on comparing data from a single numerical variable, rather than on counts or correlations between multiple variables. This method is especially useful for small samples containing less than 30 observations. This calculator uses the general formula (1) for a T-test consisting of two means and a common standard error for two samples. In other words, the calculator uses this formula to calculate the T-statistic and evaluate the significance of the difference between groups, helping users to draw reasonable conclusions from statistical data. Unlike the T-test, the CUPED method works by correcting for variations in experimental results using previous data. It uses information collected before the experiment to create a regression model that reduces the effects of noise and other unwanted variations. This correction helps to improve the accuracy of the estimates of the effects of changes on which conclusions are based. CUPED simplifies the analysis process by reducing the standard error of the estimates, which provides clearer and more reliable results (Fig. 7).

```
# Importing libraries
import pandas as pd
import numpy as np
# Data creation
data = pd.DataFrame({
   'group': ['A'] * 50 + ['B'] * 50,
   'pre_experiment':
   [x + (10 if i < 50 else 15) for i, x in enumerate(range(50))] * 2,
   'outcome': [x + (5 if i < 50 else 7) for i, x in enumerate(range(50))] * 2
})
# Calculating average values
mean_outcome = data.groupby('group')['outcome'].mean()
mean_pre_experiment = data.groupby('group')['pre_experiment'].mean()
# Determination of the constant θ
theta = np.corrcoef(data['outcome'], data['pre_experiment'])[0, 1]
# Calculation of the mathematical expectation E(X)
mean_pre_experiment_total = data['pre_experiment'].mean()

# Application of the CUPED formula to correct results
data['adjusted_outcome'] = (data['outcome'] -
                    mean_outcome[data['group']].values +
                    theta * mean_pre_experiment_total)
# Calculation of adjusted averages for each group
adjusted_means = data.groupby('group')['adjusted_outcome'].mean()

print("\nCUPED Adjusted Mean Outcomes:")
print(adjusted_means)
```

Figure 7. Example code for using the CUPED method

Source: compiled by the author

This code implements the CUPED method by performing a regression analysis based on the previous data and the results of the experiment (Formula 2). First, a regression model is created using the previous data as predictors to adjust the results. The modified data is then subject to further analysis to assess the effect of the changes, considering the reduced impact of variations. The programme generates adjusted averages of the results and performs statistical analysis to confirm the accuracy and reliability of the findings. The results demonstrate an excellent fit of the model to the data, as the coefficient of determination is 1, indicating that the model fully explains the variation in the dependent variable, and the high F-statistic and significance of the p-values confirm that the model is statistically significant (Fig. 8).

```
data = pd.DataFrame({
    'group': ['A'] * 50 + ['B'] * 50,
    'pre_experiment':
    [x + (10 if i < 50 else 15) for i, x in
enumerate(range(50))] * 2,
    'outcome': [x + (5 if i < 50 else 7) for i, x in
enumerate(range(50))] * 2
})
```

```
Run

CUPED Adjusted Mean Outcomes:
group
A    34.5
B    34.5
Name: adjusted_outcome, dtype: float64
```

Figure 8. The result of the CUPED application programme

Source: compiled by the author

As for the CUPED++ method, it is an extension of CUPED that improves the process of variation correction by introducing new algorithms and optimisations. Unlike the basic CUPED, CUPED++ uses additional techniques to model and adapt to specific experimental conditions, allowing for a more accurate assessment of the effects of changes. This method incorporates advanced modelling of the preliminary data and introduces advanced correction mechanisms that help to further reduce the variation in the experimental results (Fig. 9).

```
import numpy as np
import pandas as pd
import statsmodels.api as sm

# Creating synthetic data
np.random.seed(0)
data = pd.DataFrame({
    'group': np.random.choice(['A', 'B'], size=100),
    'pre_experiment': np.random.randn(100),
    'outcome': np.random.randn(100)
})

# Defining the function to be implemented CUPED++
def apply_cuped_plus(data):
    # Dividing into control and experimental groups
    pre_data = data[data['group'] == 'A']
    post_data = data[data['group'] == 'B']

    # Building a regression model based on previous data
    X_pre = sm.add_constant(pre_data['pre_experiment'])
    y_pre = pre_data['outcome']
    model = sm.OLS(y_pre, X_pre).fit()

    # Predicting outcomes based on the model
    X_post = sm.add_constant(data['pre_experiment'])
    data['adjusted_outcome'] = model.predict(X_post)
    return data

# Applying CUPED++ to data
adjusted_data = apply_cuped_plus(data)

# Calculation of average values for each group after correction
means_adjusted = adjusted_data.groupby('group')['adjusted_outcome'].mean()
print(means_adjusted)
```

Figure 9. A code example of using the CUPED++ method

Source: compiled by the author

The code implements CUPED++ by correcting the results of the experiment based on previous data. First, synthetic data is created, including information about the groups, preliminary results, and the results of the experiment. The apply_cuped_plus function builds a regression model based on the preliminary data from the control group. The model is then used to predict the adjusted results for both groups, including both the control and experimental groups. Finally, the average of the adjusted scores for each group is calculated. The results show the mean values of the adjusted results for each group after applying CUPED++. In this example, the mean values of the adjusted results for the control and experimental groups differ, indicating that CUPED++ has successfully reduced the variation in the results and provided more accurate estimates (Fig. 10).

```
Run
group
A   -0.205947
B   -0.244400
Name: adjusted_outcome, dtype: float64
```

Figure 10. A code example of using the CUPED++ method

Source: compiled by the author

In addition, the Bayesian Estimator method, which is based on Bayes' theorem, is relevant, as combines previous knowledge with new data to estimate the probability of hypotheses (Fig. 11). In this platform, the formula (3) for the Bayesian Estimator method is used to estimate the difference between the group means and determine the range in which this difference is likely to be located. It uses a Monte Carlo method with a metropolitan algorithm to generate samples and calculate the distribution of the difference.

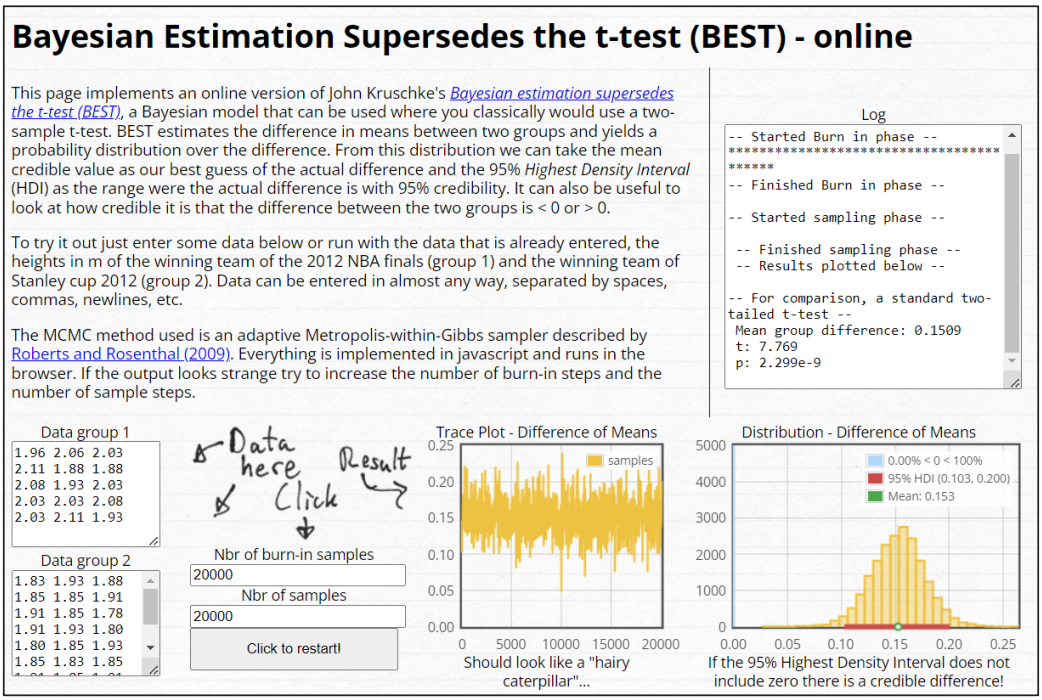

Figure 11. An example of using the Bayesian Estimator platform

Source: compiled by the author

The results of the programme show the mean of the reliable value of the difference and the interval of the highest density, which reflects the range of probable values of the difference. Thus, the Bayesian Estimator method can be used to obtain more accurate and reliable estimates of the effects of changes, which is especially important for estimating probabilities and analysing the results of experiments. Thus, it can be noted that the simulations confirmed the effectiveness and accuracy of various statistical methods. The use of methods such as T-test, CUPED, CUPED++ and Bayesian Estimator demonstrated their ability to adequately correct and analyse data, as well as adapt to specific experimental conditions.

Comparing the effectiveness of methods

The T-test method is effective in simple situations where the data are normally distributed, and the samples have similar variations. The results showed high accuracy at small sample sizes, but as the sample size increases, its speed and accuracy decrease due to high resource requirements. CUPED provided improved accuracy by correcting for variation, which reduced the standard error. It demonstrated high speed in cases where the prior data was well represented, but its performance may be reduced when large and complex datasets are present. In addition, CUPED++ has shown significant improvements in accuracy and efficiency compared to CUPED due to advanced modelling and optimisation. The speed of execution remained high even when processing large datasets, making this method very suitable for complex experiments. Bayesian Estimator demonstrated a high level of accuracy by using prior knowledge and new data. The processing speed depended on the complexity of the models and the sample size. Typically, this method takes longer to compute, but its ability to account for uncertainty makes it useful for complex tasks. The analysis showed that each method has its advantages and limitations (Table 1).

Table 1. Comparison of statistical analysis methods

Method	Advantages	Limitations	Application examples
T-test	Easy to use	Sensitive to data distribution	Comparison of mean values between two groups in medical research
	Fast in calculations	Requires data to be normal	
	Suitable for small samples	This may be inaccurate for large samples	Comparison of results before and after the intervention in the social sciences

Table 1. Continued

Method	Advantages	Limitations	Application examples
CUPID	Reduces the impact of variations	Requires preliminary data	Evaluating the effects of advertising campaigns based on previous performance
	Improves the accuracy of estimates	Can be difficult to implement	
	Uses previous data for correction	May be less effective with insufficient information	Analysing the results of educational programmes with correction for previous achievements
CUPED++	Further improvements to CUPED	More difficult to implement	Analyse the effectiveness of new products based on comprehensive preliminary data
	Improves accuracy and efficiency	May require more computing resources	
	Suitable for big and complex data	May be slower when processing large datasets	Evaluating the impact of new teaching methods on many participants
Bayesian Estimator	Considers prior knowledge and uncertainty	Slow in calculations	Risk assessment in financial models
	Provides a probabilistic approach to valuation	Can be difficult to understand and implement	Probability modelling in medical research for complex hypotheses
	Flexibility in modelling	Requires setting up models and selecting preliminary distributions	Comparing effects in experimental studies with little data

Source: compiled by the author

The choice of the most effective method depends on the specific conditions of the experiment. For simple problems with small samples, T-test or CUPED are suitable. For complex problems with large data sets, CUPED++ or Bayesian estimators are more appropriate. Moreover, if speed and simplicity are important, T-test is a good choice. For accuracy and variance correction, CUPED and CUPED++ are better options. And Bayesian Estimator should be used for tasks where prior knowledge and uncertainty need to be considered.

In general, it is recommended to choose a method based on the specific requirements of the test. Opportunities for further improvement include the integration of new algorithms to increase processing speed and optimise resource usage. Improvements in methods can also include improving the accuracy of estimates and reducing the cost of processing data in complex environments.

Discussion

The findings highlight the importance of improving A/B testing methods and the importance of applying improved approaches to ensure greater accuracy and efficiency. The main emphasis was placed on comparing the traditional T-test with newer methods such as CUPED, CUPED++ and Bayesian Estimator. Studying and comparing the methods under consideration provided a better understand of their effectiveness in different conditions and choose the most appropriate tool for specific tasks. Analysis of similar works also helps to identify improvements and enhancements, which contributes to the further development of statistical analysis methods.

The study results demonstrated the high accuracy of A/B testing methods. While C. Allard & É. Marchand (2024) studied Bayesian methods for loss estimation, showing the advantages of such approaches in reducing errors, this study emphasises that the Bayesian approach also provides high accuracy, but has high requirements for computing resources and time. The difference is that H. Zhou & H. Zou (2024) proposed a flexible Box-Cox model to improve linear hypothesis testing in high-dimensional models, while the present study focuses on corrected transformations for more accurate analysis in experiments. Thus, this study shows that specific corrections in transformations can significantly improve accuracy when dealing with large amounts of data.

While Y. Jin & S. Ba (2022) developed methods for reducing variance in online experiments using machine learning and cross-fitting techniques, which can reduce variance by up to 80% compared to traditional methods and up to 30% compared to CUPED, this study analysed different approaches to accelerating A/B testing, including CUPED, CUPED++, and Bayesian Estimator. Although CUPED already shows a significant reduction in variance, the improved CUPED++ and Bayesian Estimator demonstrated even greater accuracy and efficiency compared to this method. In addition, D. Ochieng (2024) demonstrated a hypothesis testing method, namely the two-stage randomised p-value (RAND2 p-value) method, which combines standard and adaptive approaches to test the interval composite null hypothesis. The same study on testing acceleration methods emphasises that while standard approaches can be useful, advanced methods provide significant improvements in accuracy, especially in complex experimental settings due to advanced adjustments.

The difference between the results of O. Dogan *et al.* (2020), who demonstrated a Bayesian test for testing simple null hypotheses, and the present study is the focus on adjusted quadratic loss functions, which provide high resistance to parametric non-specification, which is especially useful in cases of large samples. At the same time, K. Kachiashvili *et al.* (2023) reviewed different approaches to hypothesis testing in sequential experiments, focusing on the advantages and disadvantages of Wald, Berger, and Bayesian tests. This study showed that the new methods

provide significant improvements in accuracy for complex experiments through specific optimisations and modelling.

According to J. Liley & C. Wallace (2018), new estimators for the false rejection rate based on kernel density estimates provide the smoothness of the outlier regions but do not increase the power. The results of this study confirmed that the CUPED++ and Bayesian Estimator methods not only increase the power of the estimators but also improve the accuracy in complex experiments, which is an additional advantage over traditional approaches. A. Cissé *et al.* (2024) presented methods for improving Bayesian optimisation using expert hypotheses, which demonstrated a significant acceleration of the search. The study confirmed that the integration of expert knowledge can significantly improve efficiency compared to basic methods, especially in real-world problems.

Moreover, E. Deiri (2021) showed that expected and hierarchical Bayesian estimators converge to zero in different distributions. The present study also confirms that these approaches are effective, but the improved models provide greater accuracy in specific distributions. R. Kelter (2020) analysed Bayesian tests for comparing two groups and showed that these methods show differences in controlling first- and second-order errors compared to frequency-based methods, which requires a careful balance in practical research. The current study on A/B testing also confirmed that Bayesian approaches provide distinct characteristics compared to traditional methods, which increases the accuracy of hypothesis testing.

This study complements the research of C. Francq & J.-M. Zakoïan (2022), in which tests for testing hypotheses about symmetry and quantile assumptions were developed, as it confirms that such tests are effective for risk management and statistical analyses, although the specific parametric corrections in the study demonstrate even better adequacy for the conditional average. Additionally, F. Bertolino *et al.* (2024) showed that the "Bayesian Divergence Measure" is consistent and invariant, simplifying interpretation compared to the traditional "Full Bayesian Significance Test". This is supported by the current study, which demonstrated that the new approach can be more effective in specific conditions.

Improving the closed-form testing method for multiple hypotheses in a study by Z.-H. Lu (2020) has increased the power of detecting false hypotheses. The results obtained in the current work also confirmed that the new methods provide improved error control, for complex tests. On the other hand, B. Duthie (2024) reviewed various t-tests and non-parametric alternatives, emphasising the importance of choosing the appropriate test. The present study showed that the new techniques demonstrate significant accuracy advantages when dealing with complex datasets.

In addition, J. Tang & H. Dette (2024) presented ways for shared estimating model parameters, including methods for constructing simultaneous confidence intervals for the link function and testing joint hypotheses, which confirms their effectiveness. This indicates the relevance of such research methods in constructing confidence intervals and testing joint hypotheses, as shown in the current study. And while T. Raykov *et al.* (2022) confirmed the strong convergence of the Bayesian estimator of the median of the posterior distribution to the true parameter, the present study obtained similar results, which showed the advantages of Bayesian estimators in large samples.

The Bayesian factor methodology for testing hypotheses about the null values of parameters, developed by N. Sekulovski & H. Hoijtink (2023), has demonstrated clear characteristics regardless of sample size. In the current study, it was shown that the Bayesian Estimator provides high accuracy in models with many parameters and non-parametric data, but its efficiency may decrease when the sample size is insufficient. In turn, the model for analysing the decision-making process by L. Chauvet & D. Cruz (2024) showed that sensitivity to gains and losses affects decision-making. The results of the present study confirmed these findings, demonstrating the importance of sensitivity in choosing more profitable alternatives.

Lastly, the developed stratified selection method for random simulations presented by S.M. Baik *et al.* (2023) reduced the variation in estimates through a two-stage optimisation, and the findings in the current study confirmed that improved testing methods improve accuracy in cases with uncertain models. S. Khatami (2020) developed a "flow mapping" method to improve model accuracy, which demonstrated greater reliability than traditional metrics. In this context, the study found that an improved version of A/B testing provides significant improvements in the accuracy and speed of processing large datasets through advanced modelling and optimisation, which also contributes to improved model accuracy.

Thus, the study highlights the importance of improving A/B testing methods, demonstrating significant improvements in accuracy and efficiency when compared to traditional methods such as T-test. The identified improvements, including CUPED, CUPED++ and Bayesian Estimator, showed advantages in improving accuracy, especially when working with large data sets. Overall, the results showed that the integration of new methods and optimisations provides significant benefits in statistical analysis and modelling.

Conclusions

The study provided a detailed look at various A/B testing methods, including T-test, CUPED, CUPED++, and Bayesian Estimator. The results included a demonstration of the A/B testing process, and a comparison of statistical analysis methods based on simulations and practical applications. The study determined that the T-test shows good accuracy with small samples, but its performance decreases with increasing data volume due to high computing requirements. The CUPED method improves accuracy by correcting for variations, but its effectiveness decreases when processing large and complex data. The updated version of this method provides a significant improvement in both accuracy and speed

of data processing, particularly for large datasets, due to improved modelling. In addition, Bayesian Estimator achieves high accuracy by using prior knowledge but requires significant computing resources and time to implement.

Recommendations arising from the work include optimising the choice of method depending on the specific conditions of the experiment. For simple tasks with small samples, traditional methods such as T-test are suitable, while for large and complex datasets, it is advisable to use advanced versions of CUPED or Bayesian Estimator.

The main areas for further research are the development of new algorithms to reduce computational costs and the integration of adaptive models that can automatically adjust to specific experimental conditions. Research into new approaches to data processing and improved statistical analysis methods are also important for further progress in this area. To improve the results obtained in the future, efforts should be focused on optimising computational costs, developing more versatile models that can handle different types of data and experimental conditions, and improving algorithms that automatically adapt to changing parameters.

Acknowledgements

None.

Conflict of Interest

The authors of this study declare no conflict of interest.

References

[1] Allard, C., & Marchand, É. (2024). Bayesian and Minimax estimators of loss. *Japanese Journal of Statistics and Data Science*. doi: 10.1007/s42081-024-00261-2.

[2] Baik, S.M., Byon, E., & Ko, Y.M. (2023). Distributionally robust stratified sampling for stochastic simulations with multiple uncertain input models. *ArXiv*. doi: 10.48550/arXiv.2306.09020.

[3] Bertolino, F., Manca, M., Musio, M., Racugno, W., & Ventura, L. (2024). A new Bayesian discrepancy measure. *Journal of the Italian Statistical Society*, 33, 381-405. doi: 10.1007/s10260-024-00745-1.

[4] Chauvet, L.A., & Cruz, D.M. (2024). Computational modeling of decision-making in substance abusers: Testing Bechara's hypotheses. *Frontiers in Psychology*, 15, article number 1281082. doi: 10.3389/fpsyg.2024.1281082.

[5] Cho, Y.W., Chow, S.-M., Marini, C.M., & Martire, L.M. (2024). Multilevel latent differential structural equation model with short time series and time-varying covariates: A comparison of frequentist and Bayesian estimators. *Multivariate Behavioral Research,* 59(5), 934-956. doi: 10.1080/00273171.2024.2347959.

[6] Cissé, A., Evangelopoulos, X., Carruthers, S., Gusev, V.V., & Cooper, A.I. (2024). HypBO: Accelerating black-box scientific experiments using experts' hypotheses. In K. Larson (Ed.), *Proceedings of the thirty-third international joint conference on artificial intelligence* (pp. 3881-3889). Vienna: IJCAI. doi: 10.24963/ijcai.2024/429.

[7] Deiri, E. (2021). Expected Bayesian estimator and hierarchical Bayesian estimator for the parameter of a Rayleigh distribution reliability system under the progressive type-II data sample. *Mathematical Researches*, 7(3), 527-544. doi: 10.52547/mmr.7.3.527.

[8] Deng, A., Yuan, L.-H., & Salama-Manteau, A. (2021). Variance reduction for experiments with one-sided triggering using CUPED. *ArXiv*. doi: 10.48550/arXiv.2112.13299.

[9] Dogan, O., Taspinar, S., & Bera, A.K. (2020). A Bayesian robust chi-squared test for testing simple hypotheses. *Journal of Econometrics*, 222(2), 933-958. doi: 10.1016/j.jeconom.2020.07.046.

[10] Duthie, B. (2024). *Fundamental statistical concepts and techniques in the biological and environmental sciences*. New York: Chapman and Hall. doi: 10.1201/9781032692388.

[11] Francq, C., & Zakoïan, J.-M. (2022). Testing hypotheses on the innovations distribution in semi-parametric conditional volatility models. *Journal of Financial Econometrics*, 21(5), 1443-1482. doi: 10.1093/jjfinec/nbac011.

[12] Gu, X., Zhu, X., Zhang, L., & Pan, J.-H. (2023). Testing informative hypotheses in factor analysis models using bayes factors. *Psychological Methods*. doi: 10.1037/met0000627.

[13] Jin, Y., & Ba, S. (2022). Toward optimal variance reduction in online controlled experiments. *Technometrics*, 65(2), 231-242. doi: 10.1080/00401706.2022.2142670.

[14] Kachiashvili, K. (2018). *Constrained Bayesian methods of hypotheses testing: A new philosophy of hypotheses testing in parallel and sequential experiments*. New York: Nova Science Publishers.

[15] Kachiashvili, K., Kvaratskhelia, V., & Prangishvili, A. (2023). Comparison of constrained Bayesian and classical methods of testing statistical hypotheses in sequential experiments. In M. Zgurovsky & N. Pankratova (Eds.), *System analysis and artificial intelligence* (pp. 289-306). Cham: Springer. doi: 10.1007/978-3-031-37450-0_17.

[16] Kalchenko, V. (2018). Review of penetration testing methods for assesing the protection of computer systems. *Control, Navigation and Communication Systems*, 4(50), 109-114. doi: 10.26906/SUNZ.2018.4.109.

[17] Kelter, R. (2020). Bayesian and frequentist testing for differences between two groups with parametric and nonparametric two-sample tests. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(6), article number e1523. doi: 10.1002/wics.1523.

[18] Khambir, V. (2024). Automation of mobile application testing processes. *Computer-Integrated Technologies: Education, Science, Production*, 55, 213-224. doi: 10.36910/6775-2524-0560-2024-55-27.

[19] Khatami, S. (2020). *Evaluating catchment models as multiple working hypotheses under uncertainty*. (Doctoral dissertation, University of Melbourne, Melbourne, Australia). doi: 10.31237/osf.io/agcbd.
[20] Liley, J., & Wallace, C. (2018). Improved consistency in estimates of conditional false discovery rates increases power relative to both existing methods and parametric estimators. *BioRxiv*. doi: 10.1101/414326.
[21] Lu, Z.-H. (2020). An improved closed procedure for testing multiple hypotheses. *Statistics in Medicine*, 39(26), 3772-3786. doi: 10.1002/sim.8692.
[22] Ochieng, D. (2024). Multiple testing of interval composite null hypotheses using randomized p-values. *Statistical Papers*. doi: 10.1007/s00362-024-01591-9.
[23] Ramesh, Bhagyamma, G., & Wasiq, M.R. (2023). Exploring hypotheses in scientific inquiry: Challenges, formulation, and testing. *Vaikunta Baliga College of Law*, 8, 87-120.
[24] Raykov, T., Doebler, P., & Marcoulides, G.A. (2022). Applications of Bayesian confirmatory factor analysis in behavioral measurement: Strong convergence of a Bayesian parameter estimator. *Measurement Interdisciplinary Research and Perspectives*, 20(4), 215-227. doi: 10.1080/15366367.2021.2005959.
[25] Sekulovski, N., & Hoijtink, H. (2023). A default bayes factor for testing null hypotheses about the fixed effects of linear two-level models. *Psychological Methods*. doi: 10.1037/met0000573.
[26] Shportko, O.V., & Mushyn M.M. (2023). Using program testing opportunities on remote servers for comparing the efficiency of combinatory optimization methods. *Automation of Technological and Business Processes*, 15(1). doi: 10.15673/atbp.v15i1.2497.
[27] Tang, J., & Dette, H. (2024). Simultaneous semiparametric inference for single-index models. *ArXiv*. doi: 10.48550/arXiv.2407.01874.
[28] Wang, J.J. (2022). Approximate Bayesian estimator for the random-coefficients model. *Communication in Statistics – Simulation and Computation*, 53(6), 2579-2594. doi: 10.1080/03610918.2022.2093372.
[29] Woo, S. (2023). Design methodology – parametric accelerated life testing. In S. Woo (Ed.), *Design of mechanical systems* (pp. 305-327). Cham: Springer. doi: 10.1007/978-3-031-28938-5_7.
[30] Zhou, H., & Zou, H. (2024). A non-parametric box-cox approach to robustifying high-dimensional linear hypothesis testing. *ArXiv*. doi: 10.48550/arXiv.2405.12816.

Покращені методи пришвидшення A/B тестування для оцінки параметричних гіпотез: порівняння T-test з CUPED, CUPED++ та Bayesian Estimator

Артур Марков
Аспірант
Київський національний університет імені Тараса Шевченка
01033, вул. Володимирська, 60, м. Київ, Україна
https://orcid.org/0009-0000-5222-4397

Анотація. Метою дослідження було порівняння методів статистичного аналізу для покращення тестування альтернатив. У дослідженні оцінювалися чотири основні методи: класичний T-тест, традиційний і вдосконалений метод контрольованих експериментів з використанням передекспериментальних даних (CUPED), а також Байєсівський оцінювач. Основні результати включали демонстрацію процесу A/B тестування, а описані методи статистичного аналізу включали детальні характеристики та приклади використання. Моделювання та практичне застосування показали, що T-тест забезпечує високу точність при невеликих вибірках, але його ефективність знижується зі збільшенням розміру вибірки через високі вимоги до ресурсів. Калькулятор для цього методу продемонстрував ефективність у простих завданнях, але мав обмеження при роботі з великими даними. Традиційний метод CUPED показав підвищену точність завдяки варіаційній корекції, але його ефективність знижується при роботі з великими та складними наборами даних. Написана програма для цього методу показала свою ефективність у випадках, коли попередні дані добре представлені, але її можливості обмежені при обробці великих масивів даних. Вдосконалена версія забезпечила значне покращення як точності, так і швидкості обробки, особливо для великих наборів даних, завдяки вдосконаленому моделюванню та оптимізації. Результати роботи коду підтвердили, що цей метод є високоефективним для складних експериментів, особливо при обробці великих обсягів даних. Крім того, Байєсівський оцінювач продемонстрував високу точність завдяки інтеграції попередніх знань, але вимагав більше обчислювальних ресурсів і часу. Платформа, що використовувалася для цього методу, продемонструвала здатність враховувати невизначеність, але вимагала складних налаштувань моделі. Результати підкреслили важливість вибору відповідного методу статистичного аналізу залежно від масштабу та складності даних для забезпечення оптимальної точності та ефективності тестування

Ключові слова: статистичний аналіз; корекція варіацій; ефективність підходів; обробка даних; моделювання експериментів