Hybrid approach of Hypothesis Testing to test the mean difference between two groups utilising Gaussian Distribution and Confidence Interval

Kazi Sakib Hasan

Computer Science, Brac University, Dhaka, Bangladesh

Institutional E-mail: kazi.sakib.hasan@g.bracu.ac.bd

Personal E-mail: simanto.alt@gmail.com

Hybrid approach of Hypothesis Testing to test the mean difference between two groups utilising Gaussian Distribution and Confidence Interval

This paper presents an easier and new robust method for hypothesis testing to conclude significant mean differences between two independent or paired samples using the concepts of location, variability, confidence intervals and Gaussian distribution. For hypothesis testing of two samples, t-test is widely used. Beside this, Wilcoxon signed-rank test and often permutation test is also conducted. Each of these methods have their own rigorousness and drawbacks for which general people and non-statistics students often find it hard to conduct experiments using these. To fix these issues, a new method of hypothesis testing is proposed in this paper that basically utilises the properties of normally distributed data and resampling, and is relatively easier to calculate using only pen and paper. The time complexity analysis of each program is also conducted to give a concise overview about which hypothesis testing algorithm is more efficient and faster to execute, since statisticians use a lot of software nowadays for their analytical tasks.

Keywords: Algorithms; Analysis of Designed Experiments; Bootstrapping/resampling; Nonparametric Methods; Robust Procedures

1. Introduction

The comparison of mean differences between two groups is a fundamental concept in statistical analysis and is useful in research of multiple disciplines including medicine, social science, and engineering. Core concept of these tests is to find the mean differences among two or more groups (sample data) drawn from a population that can provide valuable insights about specific effects that create variations among the groups and infer the overall status of the population they belong to. It can often happen that the mean of one group is greater than the other but there actually exists no effect between them, rather the differences in mean are observed completely by chance. This

phenomenon is referred to as statistically not significant. So, coming to a conclusion just by comparing the arithmetic mean of the groups is not sufficient, as this may lead to errors in the result. To prove whether the means are statistically significant or not, hypothesis testing is usually conducted where the null hypothesis (H_0) usually states or denotes that there is not any significant difference in the means of the groups. On the other hand, alternative hypothesis (H_1) tries to reject the statement of null hypothesis, that explicitly explains that there is significant mean difference between the groups and the phenomenon is known as statistically significant. This particular statistical analysis falls under the umbrella of "hypothesis testing for two sample mean" which has been developed by mathematicians and researchers especially since the twentieth century after William Sealy Gosset formulated t-test and t-distribution under his pen name "Student" with a purpose to find a reliable method to ensure quality control of relatively smaller sample sizes. The problem he was facing was that the sample size of lopes and berley was small and hence replacing mean and standard deviation of the sample with the population and fitting them as normal distribution was not constructing an accurate confidence interval **[2]**. The concept of normal distribution came from the analysis of errors of measurement made in astronomical observations due to defective instruments and less-skilled experimenters. In the 17th century, while Galileo plotted these errors and drew the curve, he noticed that the curve was bell-shaped (symmetrical) and the very small errors and large errors were accumulated in the left and right side of the curve respectively, positioning the most frequent errors in the middle region of the curve. Similar phenomenon was later noticed by Laplace in 1778 when he was formulating the central limit theorem, a theorem that says that even if the univariate data is not normally distributed, the means of repeated samples from that data will still be approximately normally distributed. It indirectly means that the more randomly

generated numbers are plotted, the more the curve tends to be bell-shaped. This property in this research to create synthetic normal distributed data. However, a few decades after Laplace's observation, Adrian in 1808 and Gauss in 1809 developed the formula of normal distribution by creating the probability distribution function (PDF) of the errors and the errors of Galileo were well-fitted in that formula. In the modern era of statistics, the normal distribution is referred to as Gaussian distribution too **[5]**. It Is also to be mentioned that Laplace's role in developing similar functions was also significant. For this reason, the normal distribution curve is historically known as the Laplace-Gauss distribution **[8]**. Thus, it was slowly realised by researchers that the distributions of natural phenomena are usually normally distributed. This normally distributed dataset shows a unique property that was found out by french mathematician Abraham De Moivre. His work laid the foundation of the symmetrical distribution and the concept of standard deviation that formulated the empirical rule **[3]**. The confidence interval problem that Gosset was dealing with has a very close connection with this empirical rule, since the concept of confidence interval is based on this. As mentioned before, the existing or being created data of nature are usually normally distributed according to the central limit theorem. It happens because of the "sufficiently large" sizes of the data. The more data points are randomly added to the dataset, the more the dataset tends to be normally distributed. Therefore, sample size is a crucial attribute for a dataset to be symmetrically distributed. Now, once the properties of normal distribution; the empirical rule was formulated, statisticians began to construct confidence intervals usually with 95% level to figure out the significant threshold onto which the sample mean is very much likely to fall in. To do so, the population standard deviation needed to be known. If it was unknown for specific experiments, the sample standard deviation was used instead as a close replacement of population standard deviation if the sample

was large enough and followed the normal distribution. This is where the problem began for William Gosset. Since his samples were relatively smaller in size, the confidence interval he constructed using that became very unrealistic and hence he had to find an alternative method. In his paper titled as "The Probable Error of a Mean" (1908), he derived the formula for the t-statistic, which compares the sample mean to the population mean when the population standard deviation is unknown. The t-statistic is given by: $t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$, where $\bar{X}$ is the sample mean, μ is the population mean, s is the standard deviation, and n is the sample size. Gosset also introduced the t-distribution in his paper, which is a family of distributions that vary based on the degrees of freedom (sample size). The t-distribution is particularly useful for small sample sizes where the normal distribution's assumptions may not hold. A key contribution of his research is the concept of the "probable error," which is related to what is now referred to as the standard error of the mean. He used this to construct confidence intervals for the population mean based on the sample mean and sample size. The major contribution was explaining that the t-distribution could be used to make inferences about the population mean from relatively smaller samples that are symmetrically distributed. This was a significant advancement, as previous methods were unreliable for small sample sizes. However, although Gosset's original paper did not frame it explicitly in terms of null (H_0) and alternative hypotheses (H_1), the idea was implicit. For a single sample, the test evaluates whether the sample mean differs significantly from the population mean, which can be seen as testing the null hypothesis ($H_0 : \mu=\mu_0$) against the alternative hypothesis ($H_1 : \mu \neq \mu_0$) **[9]**. Even though Student's t-test was a novel approach to test statistical significance between the means of two groups, it has several drawbacks too. One of the drawbacks is that it assumes data maintains homoscedasticity, which means that each group has equal variances. Hence, if the

assumption is not true, then there is a good probability of inflated Type I error rate. Moreover, the method is not robust to outliers and non-normality can cause critical bias in the results as well. To create a solution for the equal variances (heteroscedasticity) problem, Bernard Lewis Welch formulated a new test that can deal with unequal variances within the two groups. The test is known as Welch's t-test. Note that, even though it successfully solved the problem for unequal variances, normality has to be maintained in Welch's t-test as well. If the data is not normal, the results may become wrong. It is also not robust to outliers. Furthermore, while Welch's t-test can handle unequal variances, it still requires reasonably large sample sizes to provide reliable results. Small sample sizes can reduce the power of the test, making it difficult to detect true differences **[4]**. Student's t-test and Welch's t-test both are the examples of parametric tests for hypothesis testing where the underlying data distribution needs to be symmetrical **[7]**. This is also a limitation of these two tests, and statisticians have solutions for these too. One better solution is the Wilcoxon signed-rank test, which is a non-parametric "rank" test used to compare two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ. Unlike the paired t-test, the Wilcoxon Signed Rank Test does not assume normal distribution of the data. It is on the basis of the ranks of the differences rather than the actual differences. The null hypothesis of this test is that the median of the differences between pairs is zero ($H_0 : M_e = 0$), meaning there is no difference between the two related groups. Conversely, alternative hypothesis stands for the median of the differences between pairs is not zero ($H_1 : M_e \neq 0$), indicating a significant difference between the two related groups. This approach is quite useful when the data is non-normal and contains outliers. Nevertheless, the method is generally less powerful

than parametric tests like the paired t-test when the normality assumption is satisfied or the sample size is quite large **[10]**. Another non-parametric test is the permutation test, that could easily surpass the efficiency of most of the parametric and non-parametric tests to compare the significance difference between two groups. A permutation test is a robust and flexible non-parametric method for hypothesis testing that makes minimal assumptions about the underlying data distribution. Basically it combines the data from two groups and then creates multiple samples by randomly picking data points from the sample without replacement. This yields a permutation distribution of the statistic of interest (e.g. mean) and later compared to the previous statistic of interest. If the statistic lies well within the permutation distribution, then it is concluded that two of the samples belong to the same population and hence the group differences are not significant. Here, the null hypothesis is that both group samples are drawn from the same population ($H_0 : F = G$) , and the alternative hypothesis is that both group samples are drawn from different population ($H_1 : F \neq G$). By relying on the rearrangement of observed data, it generates an empirical distribution of the test statistic under the null hypothesis, providing an exact p-value for the observed test statistic. However, since this method needs a lot of shuffling and looping over datasets, it can be computationally intensive than any other testing described earlier and is almost impossible to conduct using pen and paper. The time complexity of the permutation test algorithm is O(n), which denotes that the time elapsed to run the test linearly increases with the sample size which can often make

complex issues for memory and runtime both for larger datasets **[1]**. To optimise all these limitations altogether, this paper tried to come up with a new hypothesis testing that is robust to outliers, sample size, data distribution, time and computational complexity utilising the properties of Gaussian distribution, resampling, and confidence intervals.

2. Methodology

As mentioned before, the goal of this research is to develop a new hybrid method for analysing the statistical significance between two independent samples, or two pair samples that is robust to outliers, data shape, sample size, and computational complexity. Traditionally, while doing these kinds of tasks, the mean differences between the samples are measured. However, the approach I was using to conclude statistical significance, avoided this step. The approach was similar to bootstrapping and permutation tests, but less lengthier than of permutation tests, and is a combination of both parametric and non-parametric tests. Shortly, the range of each data group is measured and a normal distribution is fitted with random data points within the range. The first group was renamed as "control group" and the second group was renamed as "treatment" group following the conventions of randomised controlled trials where measuring significant mean difference is an important step. However, this test not only helps in randomised control trials, but also helps for independent samples as well. After the creation of normally distributed data for both control and treatment group, a confidence interval with 95% confidence interval was constructed for the control group's new normally distributed sample. Then, the mean of the treatment group was measured and compared with the threshold of the interval. If the mean falls outside the threshold, the significance of the mean differences were concluded as true.

Null Hypothesis H_0: $T_{mean} \in C_{mean}$. Control group and treatment group samples are drawn from the same conceptual population.

Alternative Hypothesis H_1: $T_{mean} > C_{mean}$. Control group and treatment group samples are drawn from different conceptual population.

I am labelling my approach as the "Symmetrical Synthetic Test" (SST) method to refer to it throughout the whole paper for my ease. The entire algorithm and process

of SST is clearly stated below. In the Appendices section, a step-by-step example is shown to conduct this hypothesis testing without using any computational softwares.

First of all, after the data collection process the following rule that was developed by John Tukey is applied to detect outliers and they are removed from the dataset.

Low outlier threshold = Q1 - 1.5 * IQR

High outlier threshold = Q3 + 1.5 * IQR

IQR = Q3 - Q1

Here, Q1 and Q2 is respectively the median of the lower half of the dataset (25th percentile) and the upper half of the dataset (75th percentile). Any data point that is less than the low outlier threshold or more than the high outlier threshold is considered as an outlier or extreme value.

Secondly, I calculated the range for the control group and created a new dataset of sample size of 30 with random integer and float number that exists within this range using a programming software named Python. This step could be done using pen and paper too, by continuously adding data points that are closer to midpoint of the range. It would also create a symmetrically distributed data where the mean would lie very close to the midpoint of the range. The process is explicitly described in the Appendices section. However, since I did not use pen and paper in this particular step to save my time, the data that I created from randomly assigning integer and float numbers within the range could be non-normal. Even though the central limit theorem says that sample size of 30 tends to be approximately normally distributed, the distribution may not be always normally distributed due to randomness **[6]**. To solve this issue, I again calculated the arithmetic mean and standard deviation of the data and used these statistics to fit a new normally distributed data for the control group. Python can return a

perfectly normally distributed data if we pass the sample mean, standard deviation and sample size to the parameters of np.random.normal (loc, scale, size). NumPy, conventionally abbreviated as np by programmers, is a library of Python specially made for data science, “random” is a module of NumPy library, and normal() is a function belonging to the module that returns normally distributed data based on the parameters “loc”, “scale”, and “size” that respectively takes arithmetic mean, standard deviation, and sample size as arguments given by the programmer. Since, the new control data became normally distributed and it was absolutely certain, I constructed a confidence interval with 95% level. This level is considered as the significance level of hypothesis testing.

Thirdly, again I calculated the range for the treatment group and created a new dataset of sample size equal to 30 with random integer and float numbers that exist within the range. Similar to the second step, I calculated the arithmetic mean and standard deviation of the newly formed dataset and passed these arguments to NumPy’s normal() function. This time, once the new symmetrically distributed data for the treatment group is generated, instead of constructing a confidence interval, I calculated the arithmetic mean only and compared it with the previously constructed confidence for the control group. If the mean falls within the confidence interval, I concluded that the mean differences were not statistically significant. Because it proved that the samples were actually drawn from the same population, which means that if the control group was belonging to specific participants for a drug test who was given placebo and the treatment group was belonging to the participants who were given the real drug was actually from the same population. Now, if the control group and treatment group belongs to the same population, then we cannot say that the drug was indeed useful. If it was useful, the conceptual population for control group and treatment group would be

different; one specific parameter would fit for placebo group, and that same parameter with different value would fit with treatment group to describe the variability in attributes of the groups. On the other hand, if the calculated mean for the treatment group was outside the confidence interval of the control group, I concluded that there is a significant difference. It would denote that both samples are from different populations and the population mean for the treatment group is significantly higher than the control group. The concept is very easy to understand. According to the properties of Gaussian distribution, approximately 95% data points fall within 2 standard deviations from the mean. So, creating a confidence interval with 95% level means that within that interval, 95% data point of control population would fall. Now, if the calculated mean of treatment data does not fall within this range, we can clearly state that the treatment group belongs to an entirely different population whose mean is too higher than that of the control group and that the mean is an unnatural data point with respect to the control group. With 95% confidence, it can be stated that the control group and treatment group belong to different conceptual populations. This unnaturalness proves that the mean difference between the control group and treatment group is significantly different.

However, a hypothesis is only not tested with 95% significance level. It can be rather 99%, 90%, 80%, 50% or anything between 0% to 100%. To conduct hypothesis tests on a different level than 95%, the confidence interval should be constructed based on that significance level. For example, if it is necessary to conduct a hypothesis test with 90% confidence, then the researcher has to go through the table of standard deviations and its respective data point coverage, which is 90% in this context. The table is given in the supplementary materials. Then, the confidence interval should be compared with the treatment group's mean to infer a conclusion.

It was important to find a way to validate the approach of SST. Since traditional tests like Student's t-test, Wilcoxon signed-rank test, and permutation test are already known to be accurate, therefore the Symmetrical Synthetic Test approach of conducting hypothesis testing to get insights about the significant mean differences between two groups was later compared with these existing approaches. The time complexity of the three existing tests and SST is also analysed to figure out which test provides the result faster and how the runtimes vary according to the sample size. This step was done as most of the statisticians now use computational softwares to conduct hypothesis tests. The runtime analysis of the tests also shows meaningful insights about the underlying algorithms and processes of the tests. Hypothesis testing, or any sort of computer, or mathematical programs that require rigorous calculations that cannot be done by humans easily usually takes more time. One major goal of this research paper is to find an alternative of traditional hypothesis testing that can be easily calculable, understandable, predictable and applicable for general people, not for statisticians only. Note that this efficiency measurement between SST and other approaches was conducted using Cohen's d as it not only concludes statistically significant decisions, but also focuses on effect size for practical applications. **[2]**

In order to check the accuracy of SST, a Python program was written that would randomly take sample mean, and standard deviation from a list and then create two normally distributed data with given sample size using np.random.normal(). It is necessary for conducting t-test that the data should be normally distributed and the variances between two samples should be equal. Therefore, I decided to test the algorithms using symmetrically distributed data only. The program would run 10 times using a for loop, and every time it would take random mean and standard deviation to create newer samples. The sample size in each loop was programmed to be incremented

by 5 and the runtimes were being calculated and recorded in lists whenever the code blocks for Student's t-test, Wilcoxon signed-rank test, Permutation test or Symmetrical Synthetic Test were being run. The sample size was being incremented by 5 in each iteration of the for loop to analyse how each algorithm runs according to the sample size.

In each iteration of the for loop, the program was randomly creating two samples with the respective sample size. For example, in the first iteration the sample size was 5, in the second iteration it was 10, and so on and so forth. The two samples then were used to conduct hypothesis testing. Firstly, I tested the similarities between t-test and SST. To do so, I conducted a hypothesis test to find the significant mean differences between the two generated samples with 95% significance level for avoiding difficult calculations. For t-test, the condition was set in this manner: if p-value is less than 0.05, then the program would return "Reject the null hypothesis" as a string. Otherwise, it would return 'Fail to reject the null hypothesis". Similarly, if the treatment (sample 1) sample's synthetic mean exceeds the 95% confidence interval of control (sample 2) sample, then the program would return "Reject the null hypothesis", otherwise it would return 'Fail to reject the null hypothesis". Thus, in each iteration, if the program returns the same statement for both of the tests, a previously assigned variable named "accuracy" increments by 1. The maximum value of this accuracy could be 10 as the loop ran 10 times, which includes the total number of testruns. After comparing the accuracy of t-test and SST, comparisons of other tests with SST were also done. Moreover, the time complexity and runtime of every test were being measured and later plotted to gather insights about the efficiency of the test algorithms. It should be mentioned that I used Python's SciPy library's "stats" module to import the "ttest_ind(), and wilcoxon()" functions in order to conduct t-test, and Wilcoxon signed-

rank test respectively. For the permutation test, I wrote the algorithms using AI, as Python does not have any built-in function to conduct this test like it has for t-test and Wilcoxon signed-rank test. Therefore, it may happen that these programs' runtime complexity in Python will not be matched with other softwares complexity as well.

3. Results and Discussion

3.1 Accuracy of SST with other tests

The accuracy of SST was 90%, 100%, and 90% respectively for Student's t-test, Wilcoxon signed-rank test, and permutation test. In all 10 hypothesis tests, the SST and t-test along with the permutation test showed similar conclusions 9 times, and SST and Wilcoxon signed-rank test showed similar conclusions every time. It proves that the SST might conceptually lie very well with the traditional hypothesis testing algorithms. However, it was impossible to get insights about Type I and Type II errors. Moreover, the contradiction that t-test and permutation test made with SST one time, it was not understood which test was actually providing the truth conclusion. Or, it can also happen that the conclusion was due to a small error that occurred due to randomness in any of the approaches.

3.2 Efficiency comparison of SST and other tests

The following table shows the runtime efficiency of Symmetrical Synthetic Test, Student's T-test, Wilcoxon signed-rank test, and permutation test.

SST Algorithm Runtimes (s)	T-Test Algorithm Runtimes (s)
0.001451	0.006062
0.000243	0.001077

0.000211	0.00093
0.000268	0.000813
0.000224	0.000811
0.000238	0.000778
0.000228	0.00077
0.000239	0.000759
0.000238	0.000766
0.000247	0.000788

Table 3.1: Runtime comparison of Symmetrical Synthetic Test and Student's T-Test.

Using statistical software, it was found that the value of Cohen's d is -0.82, which includes a large effect size according to the magnitude. It denotes that the difference in runtimes between SST and T-test is substantial. In practical terms, this suggests that SST is significantly faster than T-test and this difference in performance is not just statistically significant, but also practically meaningful.

SST Algorithm Runtimes (s)	Wilcoxon Signed-Rank Test Algorithm Runtimes (s)
0.000408	0.001397
0.000391	0.002372
0.001841	0.001612
0.000227	0.001401
0.000246	0.000851
0.000241	0.000919
0.000325	0.000962
0.000309	0.000987
0.000249	0.001047
0.000277	0.001035

Table 3.2: Runtime comparison of Symmetrical Synthetic Test and Wilcoxon Signed-Rank Test.

From table 3.2, using statistical software, it was measured that the d value is -1.68. Similar to the previous explanation, SST is significantly faster than the Wilcoxon Signed-Rank test and is practically more applicable.

SST Algorithm Runtimes (s)	Permutation Test Algorithm Runtimes (s)
0.000753	0.220513
0.000452	0.174862
0.000464	0.172418
0.000528	0.177309
0.000319	0.176601
0.000332	0.179008
0.001704	0.196895
0.000458	0.181461
0.000391	0.18716
0.000336	0.192773

Table 3.3: Runtime comparison of Symmetrical Synthetic Test and Permutation Test.

This time, the value of Cohen's d was found to be -18.03, proving a very large effect size and statistical significance along with practical implications between the two algorithms. Practically, the SST algorithm is better than the permutation test algorithm. However, the number of permutations was set 10000 in the parameter by the AI while it was writing the code. It can be a case why d value is this much larger in this case.

Fig 3.1 depicts the overall runtime status according to the increment of sample sizes.

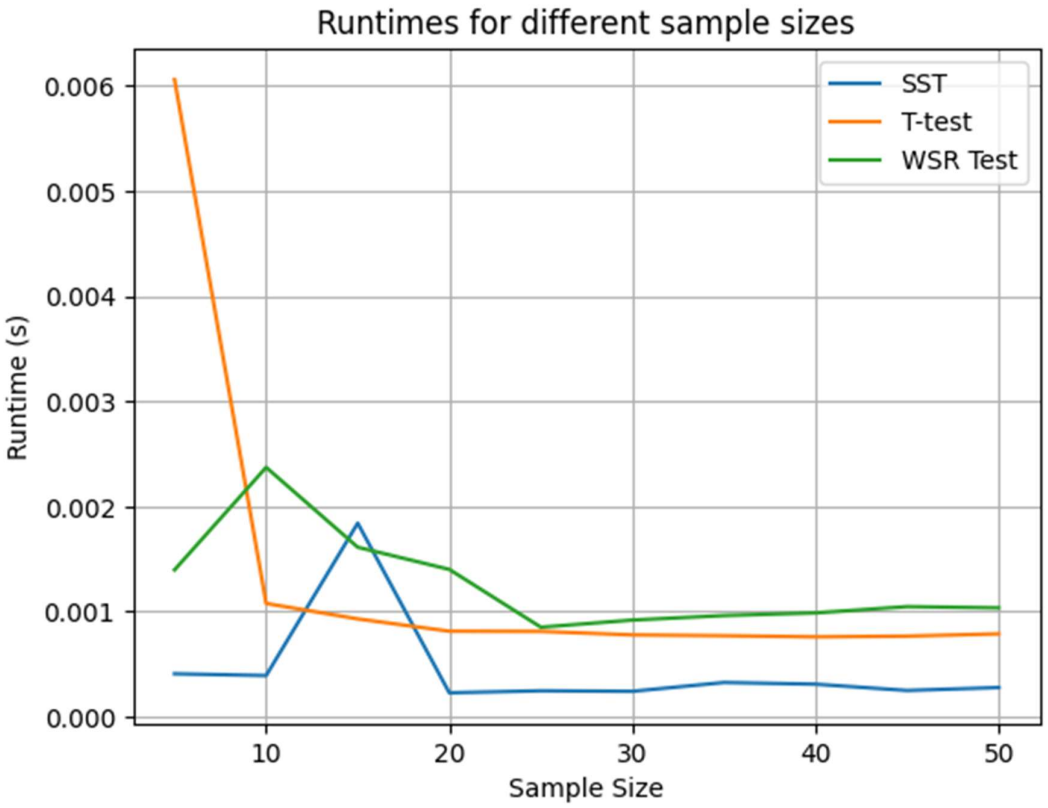

Fig 3.1: Hypothesis testing algorithm runtimes vs sample size (Permutation test excluded)

From Fig 3.1, it is noticed that the runtimes of SST was most of the time lower than the other two hypothesis testing algorithms. Surprisingly, Wilcoxon signed-rank test and SST showed a very similar plot; they were also concluding the same decision when tested for accuracy check. T-test ran slower than Wilcoxon signed-rank test and SST for the first sample size (size = 5), and then descended quickly when the sample size became 10. Since that time, its runtime continuously began to flow below the Wilcoxon signed-rank test, proving that it is a slightly better algorithm than the Wilcoxon signed-rank test. However, the permutation test plot is excluded in Fig 3.1 due to its relatively larger runtimes. Among the four algorithms, it is the most complex one. The mean runtimes of SST, T-Test, Wilcoxon signed-rank test, and permutation test are 0.0004s, 0.0013s, 0.0012s, and 0.1859s respectively.

4. Conclusion

The paper has shown an optimised robust approach to conduct hypothesis testing for comparing two means. Traditional approaches of comparing means have their own drawbacks regarding outliers, sample size, distribution shape, homoscedasticity etc. Hence, this paper hence tried to find a solution for them using the properties of Gaussian distribution, confidence intervals and synthetic resampling. Furthermore, these traditional approaches, that can be divided into popular parametric and non-parametric tests, are quite difficult to understand for general people who do not belong to the applied statistics discipline. But hypothesis testing is an important tool in the field of statistics and it provides useful insights about the data that exist in our nature. So, it is very necessary to understand this concept for most of the students or general people. In our society, almost every person knows how to calculate arithmetic mean because they can realise the importance of it. Hypothesis testing, on the other hand, is a rigorous concept and therefore people do not know anything about it even though it is a very necessary tool to understand the surroundings better. For this reason, this paper presented a new approach of conducting similar hypothesis tests like t-test, Wilcoxon signed-rank test, and parametric test just by using the relatively easier concepts of measures of central tendency and variation, confidence intervals, and normal distribution. Furthermore, today is the era of computer science and big data. Hence, statisticians prefer to use softwares a lot in their analysis project. It is essential to understand the time complexity issues of each hypothesis testing algorithm in order to cope up with unwilling errors and program crashes. Therefore, a brief overview on the runtime efficiency of different algorithms are also discussed in the paper. However, this research still has some limitations. The pen and paper method mentioned in the Appendices section might not always create normally distributed data, the data can be

skewed sometimes. Also, the contradiction in concluding a decision of the SST model with t-test and permutation test is not investigated in this paper. The accuracy was measured by comparing the algorithms 10 times only. It can happen then if the comparison was done more times then the accuracy would increase or decrease accordingly. Therefore, professional and experienced researchers in this field can conduct further research to solve these mysteries.

5. Supplementary materials

The codes that are written for conducting this research using Google Colaboratory, an online Python 3 interpreter, are given here:
https://colab.research.google.com/drive/17_O06Y0eWzIZ510s-3oJipM8UOgG6Vs8?usp=sharing
Note that, due to lack of time, not all codes were possible for me to write on. Several codes were written by AI (ChatGPT-3.5) at my prompt. This code includes the program of Cohen's d and the function for the permutation test.

The following link is the Google Spreadsheet that contains the data of the four hypothesis testing algorithms (t-test, SST, Wilcoxon signed-rank test, and permutation test) and their runtimes.
https://docs.google.com/spreadsheets/d/1dLKnvPWS19VytpH7I46nL18giSoxUVMyVMTxzduBNVg/edit?usp=sharing

6. Appendices

6.1 Appendix A

To create the confidence interval of a symmetrically distributed data with a certain level, the sample mean and respective coefficient of standard deviation to that level should be known. For example, if we are to construct the 95% confidence interval, then the interval will be (sample mean - 2 * standard deviation, sample mean + 2 * standard deviation). Here, the coefficient of standard deviation is 2 since in a normally distributed data, approximately 95% of the data points fall between 2 * standard deviation from the mean. The list of other confidence level and data coverage is given here: https://docs.google.com/spreadsheets/d/1-1VOQcc6pK_s78pPQkAysCvOqnB46k7g/edit?usp=sharing&ouid=110535718283232320699&rtpof=true&sd=true

6.2 Appendix B

The creation of normally distributed data using only pen and paper is mentioned in the paper. Here is how to do that.

Assume, a researcher is conducting a hypothesis test to figure out whether a medicine has an effect in curing a headache or not. She has the following data, which shows the maximum time of headache for 5 patients before taking the medicine and after taking the medicine.

before = [7.5, 2.75, 3, 4.5, 3.5]

after = [5.5, 1, 2.25, 2, 2.5]

Since, the researcher is trying to prove that the pain time (hour) is reduced after taking the medicines, she will (according to the research) assign “control” variable to after and “treatment” variable to before. Because the alternative hypothesis of SST can only prove that a group mean is significantly higher than the other or not, we need to convert our necessary data in such a way.

Null Hypothesis $H_0 : T_{mean} \in C_{mean}$

Alternative Hypothesis $H_1 : T_{mean} > C_{mean}$

Control = [5.5, 1, 2.25, 2, 2.5]

Treatment = [7.5, 2.75, 3, 4.5, 3.5]

Control range = (1, 5.5)

Treatment range = (3,7.5)

New_control = [1,1.25,1.5,3.35,3.25,3.5,3,3.5,3.25,3.75,4,3.75,3.25,3.25,5.5] (divide the range by 2. Most of the data points should be around that yielded value to form normally distributed data)

Standard deviation of New_control = 1.10

Mean of New_control = 3.14

95% CI = (3.14 - 2 * 1.10 , 3.14 + 2 * 1.10) = (-0.29, 5.34)

New_treatment = [3,3.25,3.5,7.5,7,5.25,5,5.5,5.25,5.75,6,6.25,5.5,5.75,5] (following the same thumb rule during the creation of New_control)

Mean of New_treatment = 5.3

Falls within the interval of New_control. Hence, we cannot reject the null hypothesis under 95% significance level. The mean differences within the two groups are not statistically significant.

Using statistical software, we will notice that the p-value for the before and after dataset is 0.2, which denotes the rejection of the null hypothesis under 95% significance level.

References

[1] A. Bruce, P. Bruce, and P. Gedeck, *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python*, O'Reilly Media, 2020.

[2] B. Illowsky and S. Dean, *Introductory Statistics*, Rice University, 2018.

[3] C. Clement, *The History of 68.2.95.4.99.7 in Statistics* (2019), Available at https://towardsdatascience.com/the-history-of-68-2-95-4-99-7-in-statistics-82fdcef0a0ea

[4] D. Kurtis, *Welch's t test is more sensitive to real world violations of distributional assumptions than student's t test but logistic regression is more robust than either* (2024), Available at https://link.springer.com/article/10.1007/s00362-024-01531-7

[5] D. Lane, *History of the Normal Distribution* (2020), Available at https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Introductory_Statistics_(Lane)/07%3A_Normal_Distribution/7.02%3A_History_of_the_Normal_Distribution

[6] D.C. Montgomery and G.C. Runger, *Applied Statistics and Probability for Engineers*, John Wiley & Sons Inc., 2011.

[7] M.J. Campbell, and T.D.V. Swinscow, *Statistics at Square One*, Wiley. 1976.

[8] S. Prokhorov, *Gauss-Laplace Distribution* (2014), Available at https://shorturl.at/jWrJD

[9] Student, *The Probable Error of a Mean* (1908), Available at https://doi.org/10.1093/biomet/6.1.1

[10] W.W. Daniel and C.L. Cross, *Biostatistics: A foundation for Analysis in the Health Sciences*, John Wiley & Sons Inc., 2020.