

Liam Jenkins 2502497

by Liam Jenkins

Submission date: 09-Jan-2026 01:57AM (UTC+0000)

Submission ID: 272621405

File name: 66128_Liam_Jenkins_Liam_Jenkins_2502497_709067_219932730.docx (174.64K)

Word count: 2148

Character count: 11213



DATA ANALYSIS AND VISUALISATION

Assignment 2

Liam Jenkins (2502497)

Contents

Introduction 2

Investigation of the Dataset 3

Hypothesis Development 6

Selection and justification of appropriate techniques 7

Application of selected techniques 8

Results and Discussion 9

 The Test Outcome..... 11

 Assumptions and interpretation 12

Conclusion..... 13

Bibliography 14

Introduction

Network latency is critical for performance in modern networks, with a direct influence on responsiveness and the quality of service. This report will be investigating how viable a newly proposed routing algorithm is by analysing its effect on the latency of the network, using measurement collected from a set of data packets from two routing conditions. Comparing an existing algorithm (Server 1) and a new routing algorithm (server 2).

The report presents the development and testing of a relevant and appropriate hypothesis, it identifies and evaluates outliers, discusses assumptions made and possible limitations and provides a recommendation based on the data, with statistical significance being tested at 0.10, 0.05 and 0.01 levels.

Investigation of the Dataset

The dataset used consist of 42 measurements of the latency in milliseconds, for data packets processed by two different servers. Each packet was processed by both servers 1 (the existing routing algorithm) and server 2 (the new routing algorithm).

The same data packets are used in both conditions therefore the observations are paired; this allows for packet specific variability to be controlled. This design strengthens confidence by reducing unexplained variation due to other factors.

Due to the paired nature of the data comparisons are made on the difference in latency between server one and server two for each packet, instead of treating the two servers as independent samples.

To investigate the data a boxplot was created to examine how the data is distributed and how variable the response time for both algorithms were.



Figure (1) showing the created boxplot.

Findings through investigation of dataset and statistics.

Before formal hypothesis testing, basic characteristics of the dataset were examined. These include the number of observations the paired structure of the dataset, and the range of values from the latency in the dataset.

The data indicates that server 2 frequently records lower latency values than server 1 for the same packet, this may initially suggest an improvement to the performance. However descriptive observations on their own are insufficient to say whether or not this is true as the difference may be due to random variation and not be statistically significant.

In addition, there is presence of unusually large latency values raising questions to reliability. The observations emphasise how important the application of an appropriate statistical test is in assessing the two algorithms.

A histogram of the differences in latency was produced to look at the distribution of the paired differences. It's also found that there was one significant outlier with one result from server 1 having a 140ms latency, for the tests I decided not to remove it but to instead keep it in, however I would also perform a test without this outlier. It seemed best to keep the outlier in place as in the real world where more packets are likely to be getting passed through the server's outlier like this one may more commonly show up throughout the usage, however, just to ensure confidence in the statistical testing it seems best to also test without the outlier.

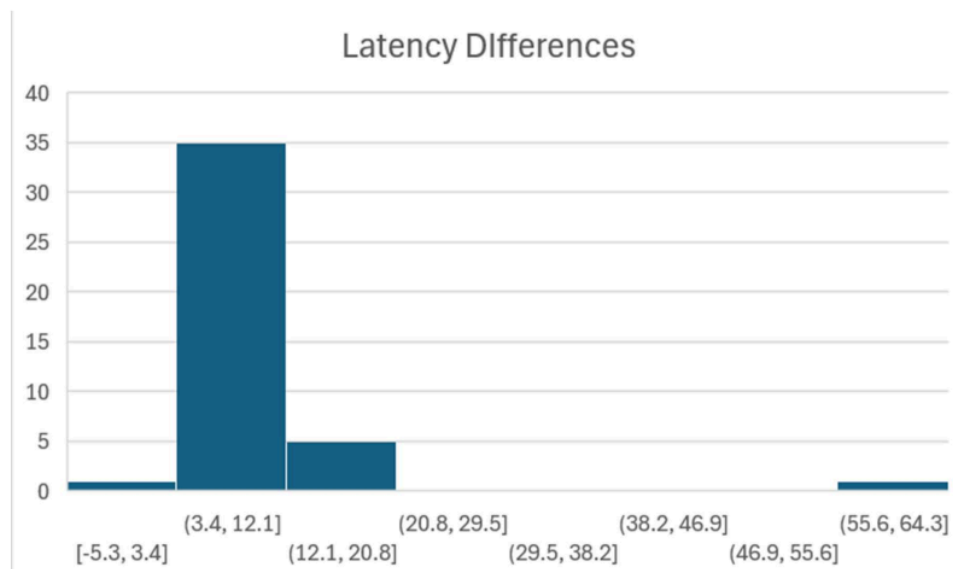


Figure (2) showing the histogram of latency differences.

Hypothesis Development

To determine whether the new routing algorithm is affecting the latency, formal statistical hypotheses were developed.

If we set “d” be equal to “Latency of server 1 – Latency of server 2” we outline the hypothesis. This would mean the null hypothesis is $H_0: \mu_d = 0$ this would mean the alternative hypothesis is $h_1: \mu_d \neq 0$.

Where μ is the mean.

With our null hypothesis being that there is no difference in the mean latency between the existing routing algorithm and the new routing algorithm and the alternative hypothesis being that there is a difference in the mean latency between the existing routing algorithm and the new routing algorithm.

Selection and justification of appropriate techniques

Due to how the given data is paired, a paired samples t-test is the best selection for statistical testing. This is appropriate due to fact we are comparing the means of two related samples, such as measurements taken from the same data packet under different conditions [1].

Using a paired test is preferable to an independent samples t-test, as it allows us to account for the variability of the packets giving more confidence in the conclusion. As latency is a continuous variable measured on an interval scale, it supports the usage of this test further.

The paired t-test assumes that that the differences between the two datasets are close to being normally distributed. Whilst it's not guaranteed that they will be exactly distributed normally, the sample size of 42 provides confidence against non-normal deviations [2].

Significance levels of 0.10, 0.05 and 0.01 were selected to allow an evaluation of the results at different levels of statistical significance, this is common practice in the field of applied statistics [3].

Application of selected techniques

To apply the selected statistical technique being the t-test, the latency for the new routing algorithm was subtracted from the latency recorded from the existing algorithm for each data packet. This produces a single set of paired latency differences.

A paired sample t-test was then performed on these differences to test whether the mean difference in latency significantly differed from zero which is our null hypothesis. The resulting test statistic and p-values were then compared to each of the significant levels in place being 0.10, 0.05 and 0.01.

The test was also repeated after the removal of the largest latency value to see the effect of extreme observations. This additional test was used to see if the presence of the outlier had a substantial effect on the outcome of the test.

Results and Discussion

Using the paired t-test we collected the t-obtained value which allows us to decide whether using the data and information we have if we accept the null hypothesis and reject the alternative hypothesis, or if we can reject the null hypothesis and accept the alternative hypothesis. The t-obtained value or the t-statistic was 8.47.

The results from the test removing the data packet sample that was an outlier gave differing yet similar results. The test statistic for the new dataset which only contained 41 samples of data was 22.12715, this is a value much higher than using the data with the outlier in it.

When combining this with a critical value, which can be obtained via stats tables by taking the number of samples in this case its 42 and subtracting one resulting in 41 gives us the degrees of freedom, this can then be used to find critical values for the different significant levels that we wish to test at being 0.10, 0.05 and 0.01. However the way I obtained the critical value was using a software package in this case Microsoft Excel, via the usage of its data analysis tools I gathered both the t-statistic and critical value, however this alone was not enough as it only gave critical values for 0.10 and 0.05 significance levels with me instead having to use a formula to generate the significance to a 0.01 significance level. In the end the critical values of 1.6828, 2.0195 and 2.7012 were given for each significance level respectively.

Since we also tested without the outlier that meant that there was a lesser number of samples, this mean subtracting 1 from 41 would be done instead resulting in a degree of freedom of 40. The critical values for the significance levels 0.10, 0.05 and 0.01 are 1.6839, 2.0211 and 2.704459 respectively.

t-Test: Paired Two Sample for Means

	<i>Variable 1</i>	<i>Variable 2</i>
Mean	88.87143	77.61905
Variance	68.88551	9.879628
Observations	42	42
Pearson Correlation	0.090467	
Hypothesized Mean Difference	0	
df	41	
t Stat	8.474639	
P(T<=t) one-tail	7.46E-11	
t Critical one-tail	1.682878	
P(T<=t) two-tail	1.49E-10	
t Critical two-tail	2.019541	

Figure (3) showing usage of t-test in excel

The Test Outcome

From the results and critical values, we obtained and observed, the outcome of the test can be determined, from the data represented and from the statistical tests, we can see that at all three significance levels the obtained t-statistic is greater than the critical value. This means that we can reject the null hypothesis and begin to accept the alternative hypothesis.

Despite the results it cannot be guaranteed that there is truth but by testing to a significance level of 0.01 we highly reduce the chance of getting a type 1 error. This means that there is only a 1% chance of us rejecting the null hypothesis when it is true, in this context that would mean concluding the new routing algorithm affects latency when it has no effect on the latency. Thanks to this there is an incredibly high amount of confidence in saying we can reject the null hypothesis and it's very likely that we have said that correctly without errors being made.

Furthermore, we can be even more confident due to the performance of the test whilst removing the outlier within the dataset, with the t-statistic being significantly more than the critical value even with the outlier removed. For these reasons it is very safe to assume we can reject the null hypothesis and accept the alternative hypothesis.

Assumptions and interpretation

To perform these statistical tests and say whether we could reject the null hypothesis a few assumptions had to be made. The first of these assumptions is that the dataset is accurate as this greatly affects the outcome of the test, we assume the data is complete as if there was more data it may be contradictory to what has been stated and we assumed that it was free from any errors.

Another assumption made is that each record is independent of one another and each datapoint does not affect the other, if this was the case the decisions made regarding the hypothesis may be incoherent. As previously mentioned for the t-test we assume that that differences should be near to normally distributed.

It's also important to note we assume that the data that was collected is completely random, and no personal bias took place when selecting the sample, as this could force results to look more favourable for one server rather than the other [4].

From the gathered results it's fair to say with a high degree of confidence that the new routing algorithm is faster and should be put into place, this conclusion has been reached due to the extensive statistical tests performed including checking for results when the outlier isn't accounted for, even with that the statistical tests showed that the new routing algorithm is a faster and better option for reducing latency. It's safe to say implementation of the new routing algorithm within their servers would provide them with reduced latency and therefore more productivity.

Conclusion

To conclude this report, we started by analysing the dataset with minimal usage of statistics, opting instead to look at the scenario and identify exactly what the dataset is trying to get across, this allowed it to be easier to plan what to do with the data for the future.

From this starting point visualisations were created to try identify patterns within the data but more importantly identify outliers. From these visualisations an outlier was identified, this meant decisions on whether to include it within the statistical tests that would be chosen. In the ending opting for two tests one with and one without the outlier.

Then the development of a suitable hypothesis for the scenario had to be created, in the end a null hypothesis and alternative hypothesis that were incredibly relevant to the scenario at hand were developed, allowing statistical testing to take place.

From the format of the data presented one main choice for the type of statistical test became apparent that being a paired t-test, this was due to the nature of comparing two means. This was the test that would give the most confidence when coming to rejecting or accepting the null hypothesis.

The chosen statistical test was then applied to yield results giving the information desired with the knowledge of the reasonable assumptions made. The statistical test allowed us to reject the null hypothesis and conclude that the new routing algorithm was faster than the old routing algorithm. This conclusion was reached to a high degree of certainty due to the testing with and without outlier and the differing significance levels tested at.

Bibliography

- [1] "Paired t-Test," JMP Statistics Knowledge Portal, JMP Statistical Discovery LLC. [Online]. Available: <https://www.jmp.com/en/statistics-knowledge-portal/t-test/paired-t-test>. Accessed: 05 Jan. 2026.
- [2] T. K. Kim and J. H. Park, "More about the basic assumptions of t-test: normality and sample size," Korean Journal of Anesthesiology, vol. 72, no. 4, pp. 331–335, Aug. 2019. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6676026/>. Accessed: 07 Jan. 2026.
- [3] S. Tenny and I. Abdelgawad, *Statistical Significance*, StatPearls [Internet]. Treasure Island, FL, USA: StatPearls Publishing; 2025 Jan–. [Online]. Available: <https://www.ncbi.nlm.nih.gov/books/NBK459346/>. Accessed: 07 Jan. 2026.
- [4] "The t-Test," *JMP Statistics Knowledge Portal*, JMP Statistical Discovery LLC. [Online]. Available: <https://www.jmp.com/en/statistics-knowledge-portal/t-test>. Accessed: 08 Jan. 2026.

FINAL GRADE

GENERAL COMMENTS

50/100

Generic Assessment Criteria	Marks available	Marks awarded
1. Engagement with Literature Skills	10	
Bibliography and referencing present and reasonable.		5
2. Knowledge and Understanding Skills	30	
Reasonable knowledge demonstrated of the outliers and discussion of descriptive stats demonstrates reasonable understanding. More detailed discussion and use of tables would have been beneficial. Some reasonable visualisations however more comparison may have been advantageous with consideration of		15

correlation and further visualisations such as scatter plot and through more detailed discussion.		
3. Cognitive and Intellectual Skills	30	
Some reasonable interpretation of results however more detailed and a little more explicit discussion would have been desirable, especially around further factors. Testing with and without outliers would have been beneficial. Is this really paired data? Good choice of one tailed test and justification of tests used.		15
4. Practical Application Skills	20	
Good structure of hypotheses with justification albeit that paired was used. some use of graphs and tables evidencing the practical work. More would have been advantageous, especially the tests with and without outliers.		12
85. Transferable Skills for Life and Professional Practice	10	

A reasonably laid out, and sectioned report. Be mindful of gen AI usage, even for structure. It would be desirable to avoid first person terminology such as 'I', 'my' and 'we' in favour of third person terminology, making the report more objective and academic.		3

PAGE 1
PAGE 2
PAGE 3
PAGE 4
PAGE 5
PAGE 6
PAGE 7
PAGE 8
PAGE 9
PAGE 10
PAGE 11
PAGE 12
PAGE 13
PAGE 14
PAGE 15