Assignement 1

Billy Bouchard[1]
[1] *M.S. computer engineering, Polytechnique Montreal*

I. Introduction

CPU times are one of the most monitored statistics of all in computer science. It gives a lot of information about how the system resources are distributed, what tends to be a performance bottleneck and what kind of requests are the most done by clients. Therefore, having a robust way of analysing cpu data is important to better understand a system. This report will focus on some basic analysis techniques of 2 different CPU time datasets. It will try to answer 4 questions :

1) Are the performance data (i.e., CPU usage) normally distributed?
2) Does there exist statistically significance difference between the performance of the two versions?
3) How large is the difference if it exits?
4) Does the performance change over time?

II. Approach

A simple analytical study using Python would be enough to answer all the questions. Therefore, the first question would be answered by doing two histograms of the TOTAL_CPU column. To answer the second question, we will need to see whether they are normally distributed or not. If they are, one can use a dependent (paired) two-sample t-test. If not, a Wilcoxon signed-rank test would be used. To answer the third question, cohen's D test can be used for normal distribution otherwise the cliff's delta would be best. Finally, the fourth question can be answered with the Mann-Kendall trend test for each different set of data.

Of all those questions we shall emit 2 null hypotheses first one of which will be that

Null Hypothesis 1: There exists no relation/correlation between the original and the new data.
Since both are data that come from a server, it makes sense that a certain correlation exists between them. The second one.

Null Hypothesis 2: Over time, the average time to answer a request will be augmented.
If a system is well-optimized, there shouldn't be any issues with a large amount of requests. Therefore, the time should stay pretty consistent.

III. Results

The first question was answered by doing a simple histogram of both data that can be seen in figure 1. Distributions seem normal at first glance, but there are strong outliers on the right side of the curve which makes it derive from a normal distribution. However, since this study does not discriminate the data, we shall not remove any outliers and consider the data as a whole.

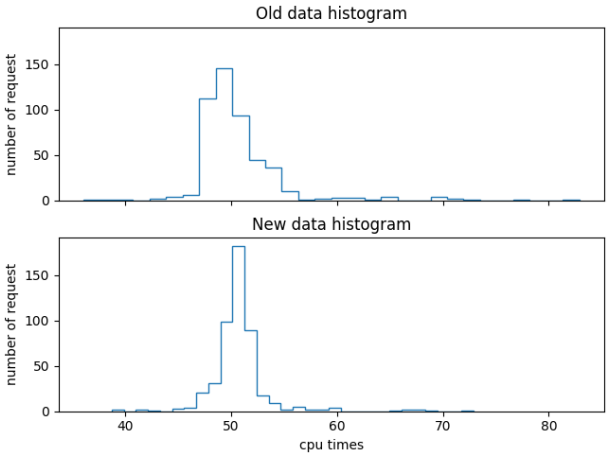

Figure 1: Histograms of new and original data

Test done	value
Wilcoxon	0.0025
T-test	0.5942
cohen's D	-0.0352
cliff's delta	-0.1976

Table I: Results of the different test done on comparing both data

Since the data distribution resembled a normal distribution, both the t-test and the Wilcoxon test have been made with the data. Results can be found in table I. This proves the fact that they do not seem to follow a normal distribution as the Wilcoxon test has a very small p-value compared to the T-test. However, from the data, it is clear that there is a correlation between both data. Thus, we can reject our null hypothesis 1, showing that there is a correlation.

To answer the third question, we also did both Cohen's D and Cliff's delta and the results can be seen in table I. However, based on the prior observations, it is probable that only Cliff's delta is accurate. The numbers show that there is only a small difference between both data sets.

Finally, to answer the fourth question, Mann Kendall test was executed on each data set to see what trend each one had. As was thought, no trend was observed even though the p-value for the original data came dangerously low. Therefore, we can reject our second null hypothesis and state that there is no trend in both data sets.

IV. Replication

All of this study, including the code and data can be replicated. These can be found at `https://github.com/34yu34/log6309e-assignement1`.

Test done	original data	new data
Mann Kendalls test	0.0596	0.2875

Table II: Trend analysis of both data set

Running the analysis.py program will show all the calculation done.

V. Conclusion

In conclusion, We were able to reject our first null hypothesis proving that there is a correlation between the original data and the new data. Moreover, the trend test confirms that there is no trend inside the data which lets us reject the second null hypothesis. However, small outliers in the data might have affected some of the tests. It would be interesting to redo the analysis of the data set once a couple of data analysis techniques have been executed on it.