MT4113 Report

"I confirm that the following report and associated code is my own work, except where clearly indicated."

Abstract

The study aims to investigate the size and power of chosen two statistical tests (one parametric – t-test and one non-parametric – Mann-Whitney U test) under different realistic scenarios within the context of collected data. The data is from the "greenhouse_gases" dataset in the dslabs package in R. The research question is "On average, is the concentration of CO_2 greater than that of N_2O?".

Four scenarios for data generation were developed. Under the scenarios, the effect of sample sizes, measurement error, standard derivations and the effect size of simulated data were examined. It found that the power of the tests increases with the increased sample sizes or the larger effect sizes. The power decreases and the size increases when standard derivations of the two populations increase. In addition, when there are large sample sizes, the power converges to 1 faster than small sample sizes.

Introduction

In order to examine the performance of the chosen statistical tests, the size and the power of the tests are considered. The chosen alpha level is 0.05. To measure the size, we figure out how often the test incorrectly concludes there is a difference between two population groups when there is no difference. To measure the power, we need to simulate a data where there is a difference, and then figure out how often the test correctly concludes there is a difference. Thus, the size and power are probabilities.

The data consisted of 100 observed concentrations of CO_2 (group 1) and 100 observed concentrations of N_2O (group 2) in ppm by volume. The data was formatted from the original dataset. Based on the research question, the chosen two statistical tests are t-test and Mann-Whitney U test. The null hypothesis is that "there is no difference between the mean of group 1 and that of group 2". The alternative hypothesis is that "the mean of group 1 is greater than that of group 2". By conducting the simulation study, we aim to answer the question: what is the size and power of these tests under realistic scenarios within the context of the collected data. The following sections in this report will first describe the methods that were used to conduct the simulation study, will then summarize the findings and will finally conclude on the performance of the tests based on our four scenarios.

Methods

After selecting the dataset and formatting the collected data, there were four steps involved in the simulation study. All analysis within each scenario was considered under the same conditions of sample sizes, means and standard derivations. Details of the simulation study method are available in the design document with specified parameter values we used based on the properties of the data.

1. Explore the data

We firstly explored the properties of the data. Then, we assessed normality and made assumptions of the two distributions. The research question and statistical tests were confirmed at this step.

2. Simulation scenarios

Scenario one

We examined the effect of sample sizes on size and power. Two cases were considered. In case 1, different numbers of concentrations of CO_2 (group 1) and N_2O (group 2) were recorded. In other words, we simulated a small sample size of group 1 ($n1$) and a large sample size of group 2 ($n2$). For example, we had $n1 = 10$ and $n2 = 50$ or 1000. In case 2, we considered the situation where there is an equal sample size between two groups; however, the sample sizes are increasing gradually. For

instance, we had $n1 = n2 = 10$ and $n1 = n2 = 1000$.

Scenario two

We explored the effect of measurement error presented in the simulated data on size and power. Since in this research, the concentration of certain gas was detected using a precise instrument, large measurement error is unlikely to occur. Thus, we considered that the reported values are rounded up to the nearest integer. We compared this scenario with scenario one under the same conditions.

Scenario three

We examined the effect of standard derivations of the simulated data on size and power. In case 1, we considered that there is a small variation in group 1 ($sd1$) and a large variation in group 2 ($sd2$). For example, we had $sd1 = 1$ and $sd2 = 1$ or 50. Then, in case 2, the standard derivations of the two population groups are the same; however, they are increasing gradually. For instance, we had $sd1 = sd2 = 10$ and $sd1 = sd2 = 50$.

Scenario four

The size of the test can be calculated when the effect size is zero. Since we considered the effect size of simulated data, we could only examine the power. Considering different sample sizes ($n1 = n2 = n$), we explored the power as effect size increases. We further analysed the impact of small effect size on sample sizes in order to achieve the power of 1 by combining scenario one and four together.

3. Generate and analyse simulated data

The main function which was used to conduct simulation and calculate the power and size of the two tests under each scenario was designed in this step. The function contained three essential sections. Firstly, we generated simulated data under four scenarios. Then, we applied the two chosen statistical tests to the simulated data and got p-value. Next, we repeated the above processes 1000 times and got 1000 p-values. Finally, we calculated the size and power of these tests under each of the scenarios.

4. Analyse size and power under each scenario

We analysed size and power by specifying the inputs of the main function under each scenario. Most importantly, the effect size should be 0 to calculate the size. To visualize the performance of the power and size of chosen tests under each scenario efficiently, we constructed functions for graphing.

Results and discussion

1. Explore the data

By conducting a two-sample t-test with one-sided alternative, we found that the p-value is less than 0.01. There is a strong evidence that the mean of the concentration of CO_2 is greater than that of N_2O. We assessed normality by constructing a Q-Q plot, which indicated that there are some outliers in the data. The outliers may be better to be removed before computing the sample means which will be used in the simulation. We could further conduct a simulation study to determine what percentage of the data to be removed. However, this is not the main purpose of this study. We also considered that the original distributions of the two population groups may be of a non-standard form and there is not enough data to examine it. Thus, we assumed that the samples are from two normal distributions with different variances according to the properties when the sample size is reasonably large.

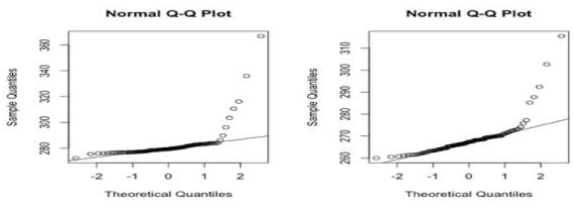

Figure 1: Normal Q-Q plot for group 1 - CO_2 (left) and group 2 - N_2O (right).

Gas	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Standard derivation
CO_2	272.3	277.6	279.2	282.1	282.2	366.8	11.8
N_2O	260.0	264.5	267.4	268.5	269.7	315.4	7.7

Table 1: The summary statistics of the collected data (group 1 – CO_2; group 2 – N_2O).

2. Analyse size and power under each scenario

Scenario one

In case 1, as the difference between two sample sizes gets larger, the power of t-test and Mann-Whitney U test stays around 0.975 and 0.99 respectively, with more variation from the t-test. The size of the t-test appears to stay around 0.05 with some randomness. However, the size of the Mann-Whitney U test appears to increase from 0.06. The results seem to meet our expectation that unequal sample sizes may affect the size of the test and cause the loss of power. In case 2, we expected that the size and power of both tests tend towards 0.05 and 1 respectively, with the increased equal sample sizes. However, the size of both tests seems to stay around 0.05 with some randomness. This is likely due to the consistent standard derivations used in the analysis as sample sizes increase. Nevertheless, the power of the tests increases and converges to 1 rapidly.

The Mann-Whitney U test generally performed better than the t-test. It may be because the variances between the simulated groups are unequal, and the parametric tests tend to be less powerful if the assumption (i.i.d) is not met. Further investigation of the original data's distributions could be conducted in order to decide on the best choice of statistical tests. However, this is not a major concern. Considering that they may not be i.i.d and the true distributions may be intricate to get, we suggest using non-parametric tests to draw inferences from future collected samples.

Test	Simulated sample size of group 1 (CO_2)	Simulated sample size of group 2 (N_2O)	Size of the test	Power of the test
t-test	10	50	0.056	0.961
	10	1000	0.063	0.980
Mann-Whitney U test	10	50	0.082	0.970
	10	1000	0.091	0.993

Table 2: The summary of the size and power of t-test and Mann-Whitney U test under scenario one case one.

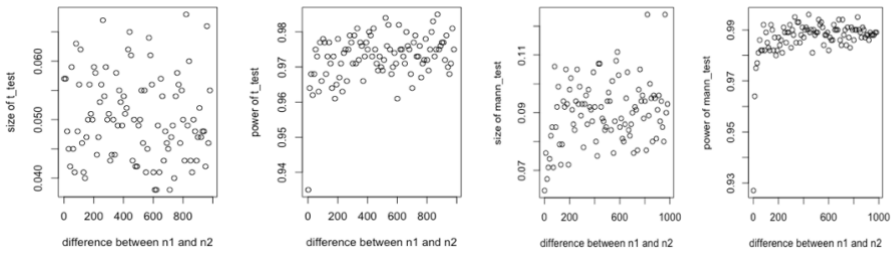

Figure 2: Plots of size and power of t-test (left) and Mann-Whitney U test (right) when difference between n1 and n2 is increasing ($n1 = 10, n2 =$ from 10 to 1000, $sd1 = 11, sd2 = 7,\ increment = width =\ 10$).

Test	Simulated sample size of group 1 (CO_2)	Simulated sample size of group 2 (N_2O)	Size of the test	Power of the test
t-test	10	10	0.078	0.935
	1000	1000	0.055	1.000
Mann-Whitney U test	10	10	0.069	0.910
	1000	1000	0.060	1.000

Table 3: The summary of the size and power of t-test and Mann-Whitney U test under scenario one case two.

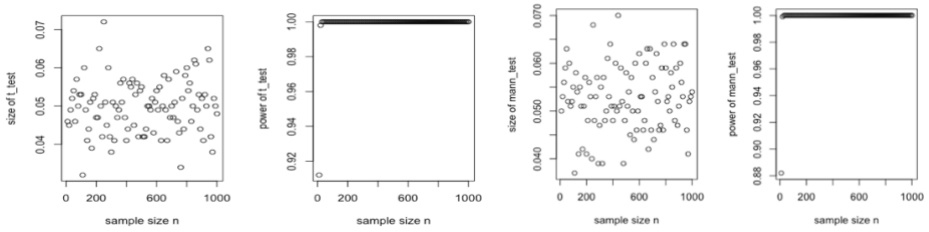

Figure 3: Plots of size and power of t-test (left) and Mann-Whitney U test (right) when n1 and n2 is increasing ($n1 = n2 = 10$ to 1000, $sd1 = 11,\ sd2 = 7,$ $increment = width = 10$).

Scenario two

There is no significant impact on the size and power of the tests as the measurement error appears. This is likely because the simulated data rounded to nearest integer is only 1 decimal place differed from the unrounded one. As the nature of this study, the concentration of certain gas was detected using a precise instrument. Thus, it is likely to be an unrealistic scenario that large measurement error such as rounded up to the nearest 5 or 10 unit occurs. However, we expect that the real effect will not be seen until the large scale of measurement error presented if the extreme case does occur. Further investigation could be conducted; however, this is not a major concern for this study.

Test	With measurement error		Without measurement error	
	Size of the test	Power of the test	Size of the test	Power of the test
t-test	0.050	1.000	0.053	1.000
Mann-Whitney U test	0.055	1.000	0.058	1.000

Table 4: The summary of the size and power of t-test and Mann-Whitney U test under scenario two.

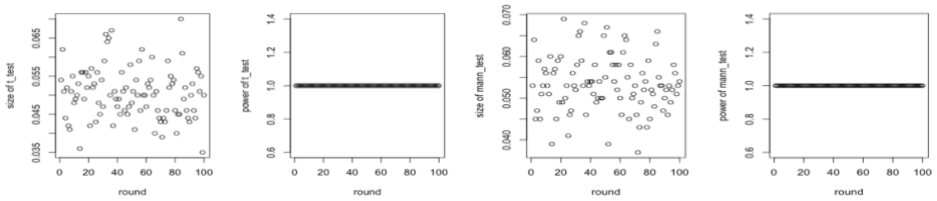

Figure 4: Plots of size and power of t-test (left) and Mann-Whitney U test (right) with measurement error ($round = 100, n1 = n2 = 100,\ sd1 = 11,\ sd2 = 7,\ increment = width = 1$).

Scenario three

In case 1, as the difference between standard derivations gets larger, the power of both tests decrease from 1 significantly. The size of the t-test stays around 0.05 with some randomness. However, there appears to be a substantial increase from 0.05 in the size of the Mann-Whitney U test, which matches our expectation. Thus, the Mann-Whitney U test seems to perform better than the t-test with the same reason stated in scenario one. In case 2 of the increased equal standard derivations within two simulated groups, as we expected, the size of both tests stays around 0.05 with some randomness, and the power of them decreases from 1 rapidly. Results from both cases suggested that large variation within the data is likely to affect statistical power and the size of the tests.

Test	Standard derivation of simulated group 1 (CO_2)	Standard derivation of simulated group 2 (N_2O)	Size of the test	Power of the test
t-test	1	10	0.038	1.000
	1	50	0.049	0.845
Mann-Whitney U test	1	10	0.078	1.000
	1	50	0.089	0.801

Table 5: The summary of the size and power of t-test and Mann-Whitney U test under scenario three case one.

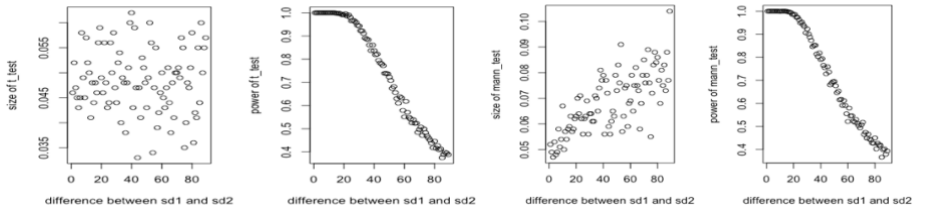

Figure 5: Plots of size and power of t-test (left) and Mann-Whitney U test (right) when difference between sd1 and $sd2$ is increasing ($sd1 = 11, sd2 =$ from 12 to 100, $n1 = n2 = 100, increment = width = 1$).

Test	Standard derivation of simulated group 1 (CO_2)	Standard derivation of simulated group 2 (N_2O)	Size of the test	Power of the test
t-test	10	10	0.053	1.000
	50	50	0.056	0.601
Mann-Whitney U test	10	10	0.047	1.000
	50	50	0.054	0.572

Table 6: The summary of the size and power of t-test and Mann-Whitney U test under scenario three case two.

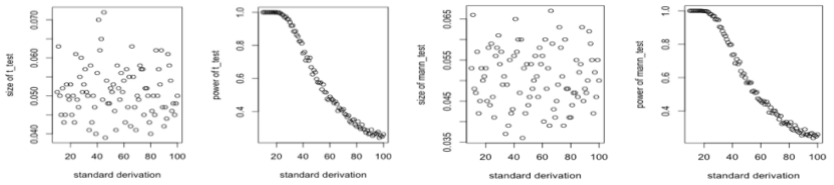

Figure 6: Plots of size and power of t-test (left) and Mann-Whitney U test (right) when $sd1$ and $sd2$ is increasing ($sd1 = sd2 =$ from 10 to 100, $n1 = n2 = 100$, $increment = \ width = 1$).

Scenario four

Given the same conditions of standard derivations ($sd1 = 11$ and $sd2 = 7$) and sample sizes ($n = 10$ or 100 or 1000), the results showed that the power of both tests increases to 1 as effect size increases, which met our expectation. Furthermore, the power of the tests converges to 1 faster as $n1$ and $n2$ increased to 1000. With further investigation of scenario one and four, given an effect size of 0.2 with $sd1 = 11$ and $sd2 = 7$, we found that 10000 samples in each group can only achieve the power of around 0.45 using both tests. Thus, we modified standard derivation of both groups to be 1 and examined the impact of small effect size on sample sizes in order to achieve the power of 1. It showed that around 800 samples per group are needed when the effect size is 0.2 for both tests. We need around 200 samples of each group to get a power of 1 when the effect size is 0.5 and 0.8. It matches our expectation that we need more samples to gain higher statistical power as effect size is small (less than 1). It also indicated that large variation could significantly cause a loss of power.

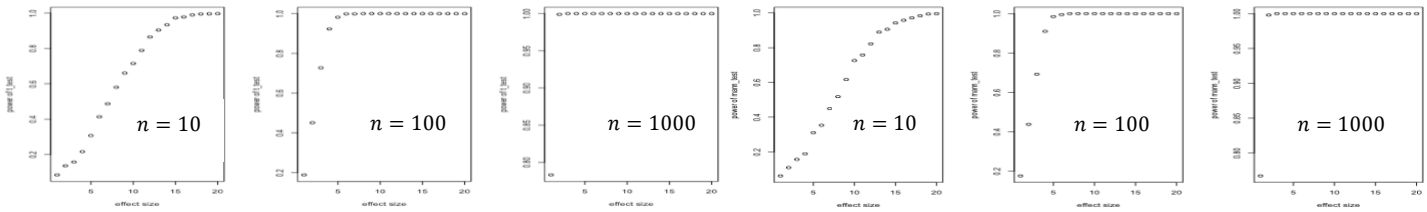

Figure 7: Plots of power of t-test (left 3) and Mann-Whitney U test (right 3) as effect size is increasing from 1 to 20 given increasing sample size $n1 = n2 = 10$, 100, 1000 from left to right ($sd1 = 11, sd2 = 7, increment = width = 1$)

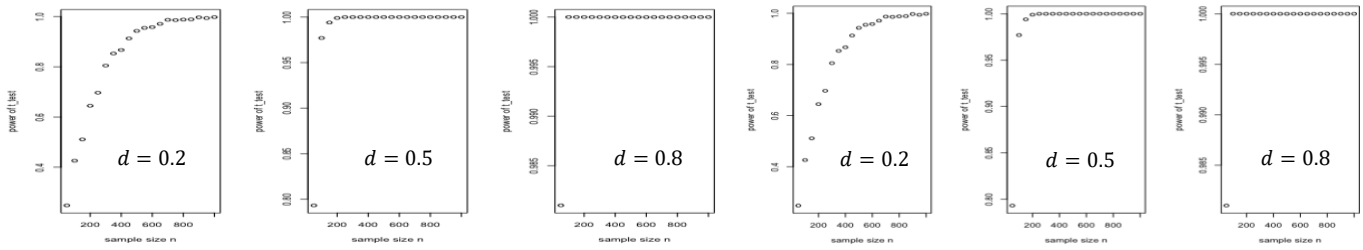

Figure 8: Plots of power of t-test (left 3) and Mann-Whitney U test (right 3) as sample sizes ($n1 = n2 = n$) are increasing from 50 to 1000 ($increment = width = 50$) given effect sizes $d = 0.2$, 0.5, 0.8 from left to right ($sd1 = sd2 = 1$)

Conclusions

To conclude, we considered the impact of sample sizes, measurement error, standard derivations and effect sizes on the size or power of one parametric test (t-test) and one non-parametric test (Mann-Whitney U test). Results have shown that unequal sample sizes and large standard derivations significantly influenced the size and power. Higher power was presented with larger effect size or larger equal sample sizes within two population groups. When effect size is small and standard derivation is large, large sample sizes are required to get higher power. Measurement error in this study had a small impact. Moreover, the size seems to fluctuate around 0.05 with consistent variations in most scenarios. Further analysis of the variations could be conducted. The Mann-Whitney U test seems to perform better generally. In addition, there are some outliers and skewness in the original data, and future samples potentially have similar trend based on the nature of the selected dataset. Thus, we suggest using non-parametric tests to draw inferences for future samples since it requires fewer assumptions and appears to have a higher chance of detecting an effect based on our scenarios.