University of St Andrews
FOUNDED 1413

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

1. Abstract

The size and power of a test can be indicators of how well the test performs. We want the power to be as higher as possible and the size as lower as possible. This report investigates how do the size and power change under different scenarios by using Monte Carlo simulation. We first generated new data sets with different properties by changing the sample size, effect size and number of decimal places. Then we applied t-test and Wilcoxon rank-sum test to each data set and got p-values from test results. Finally, we calculated sizes and powers under different scenarios and drew some conclusions. We fond that the sample size has positive relationships with sizes as well as powers of both tests. In addition, the powers of both tests can be 1 when the sample size is big enough such as 1000. Moreover, the number of decimal places has little impact on the sizes of two tests, while the effect size significantly improves powers of both tests.

2. Introduction

Monte Carlo simulation is widely used in many scientific fields such as statistics. Mahadevan (1997) explains Monte Carlo simulation as “a numerical experimentation technique to obtain the statistics of the output variables of a system computational model, given the statistics of the input variables”(p.123). In this report, we apply this technique to the US gun murder data from the dslabs package (Irizarry, 2018). The research question is whether the rate of murders in South region has significant difference with that rate in North Central. To investigate this question, one parametric test which is the t-test and one non-parametric test which is the Wilcoxon rank-sum test (also called Mann-Whitney U test) are chosen to be performed. However, answering the research question is not the aim of this report. Instead, by performing Monte Carlo simulation, this report shows how do the size and power for each statistical test change under different scenarios.

This report will first choose several scenarios for data simulation, then generate data according to the different scenarios. Subsequently, t-test and Wilcoxon rank-sum test will be executed to get p-values. After that, the size and power will be calculated for each scenario. Finally, the results will be interpreted.

3. Methods

3a) Preliminary data exploration

The data set used in this report contains the number of gun murders and the population for each state in the US in 2010. We calculated the rate of gun murders for each state by dividing the number of gun murders by the population. In addition, the rates have been multiplied by one million just to make them clearer. There are 51 states that can be grouped by four regions, which are Northeast, South, North Central and West. The Figure1 was plotted by ggplot2 package (Wickham, 2016), which shows an overall view of the differences between regions. The data in Table1 were produced with the help of dplyr package (Wickham et al., 2018), which give the mean and 95% standard deviation for each region. According to the plot and table below, the South area seems to have the highest gun murder rate, while the differences between other regions seem to be insignificant. Among these regions, we are particularly interested in whether there is significant difference between south area and north area. Therefore, data about South and North Central are chosen to be analysed in the later sections.

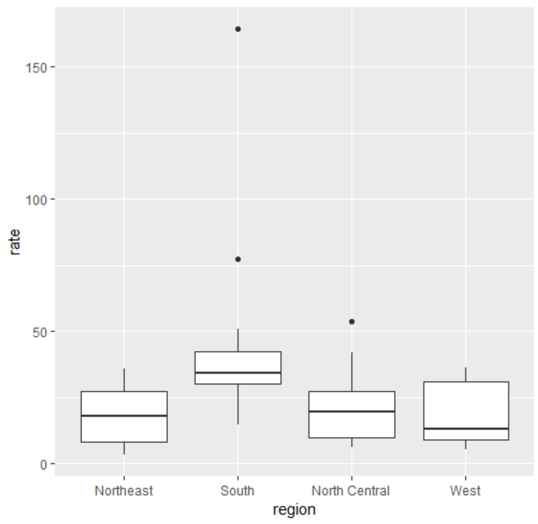

Figure1. Rate Distribution Between Different Regions

Table1. Mean and Standard Deviation of Rate for Each Region

Region	Mean	Standard Deviation
Northeast	18.475	11.741
South	44.170	33.729
North Central	21.820	14.387
West	18.334	11.722

3b) Data simulation

Three elements are chosen to develop different scenarios, which are sample size, effect size and number of decimal places. First, we controlled the effect size as zero and varied the sample size and the number of decimal places. There are three levels for the sample size, which are 10, 100 and 1000. There are also three levels for the decimal place, which are -1, 0 and 2. Therefore, there are nine scenarios with different sample sizes or decimal places in total. According to these scenarios, we used the rnorm function and round function with different values for arguments to simulate data sets with different properties. Similarly, we then generated another nine data sets by holding the value of decimal place as two and changing the value of sample size (10, 100 and 1000) and effect size (10, 15 and 20). The process of data simulation was repeated 1000 times for each scenario in order to compute sizes and powers in the later section.

3c) Size and Power Calculation

We used rnorm function to generate our new data. Therefore, the data should be normal distributed and meet all assumptions of t-test as well as Wilcoxon rank-sum test, which means these statistical tests are appropriate for our simulated data. For each simulated data set, the parametric t-test and non-parametric Wilcoxon rank-sum test were performed, then p-values are given by test results. Therefore, there are 1000 p-values from each test for each scenario. In this study, we chose 0.05 as the threshold. By calculating the proportion of time that p-values are less than 0.05, we got the size as well as power. When the effect size is zero, the proportion is the size. On the contrary, when the effect size is not equal to zero, the proportion is the power.

4. Results

The Figure2 and Figure3 below are plotted by ggplot2 package (Wickham, 2016), which illustrate the sizes and powers respectively. According to Figure2, there is significant difference between sample size groups in both tests. Furthermore, the relationships between the sizes of both tests and the sample size are positive. In other words, the probability of we incorrectly rejecting the null hypothesis increases when the sample size increases. However, the number of decimal places seems to be a less important factor. The size decreases a little when the sample size is 10, and slightly increases when the sample is 100 or 1000, but these differences are much smaller than the differences caused by sample sizes.

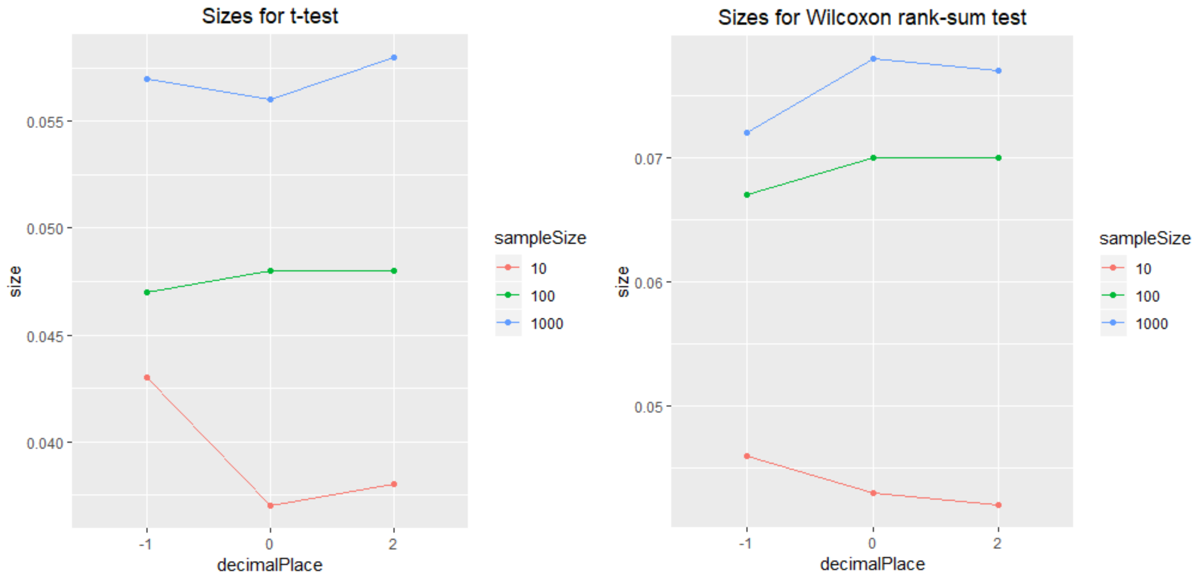

Figure2. Sizes for t-test and Wilcoxon rank-sum test under different scenarios

According to Figure3, the figures of powers for t-test and Wilcoxon rank-sum test are very similar. Like the size, the power also has a positive relationship with the sample size. Moreover, the bigger effect size also helps to improve the power of test. When the sample size is big enough such as 1000, the power reaches 1 regardless of the effect size, which means we can correctly reject the null hypothesis all the time. Besides, the gap between sample size 10 group and sample size 100 group is particularly big. The power gets close to 1 when the sample size is 100 and the effect size is 15 or 20. However, if the sample size is 10, the power is only about 0.4 even when the effect size is 20.

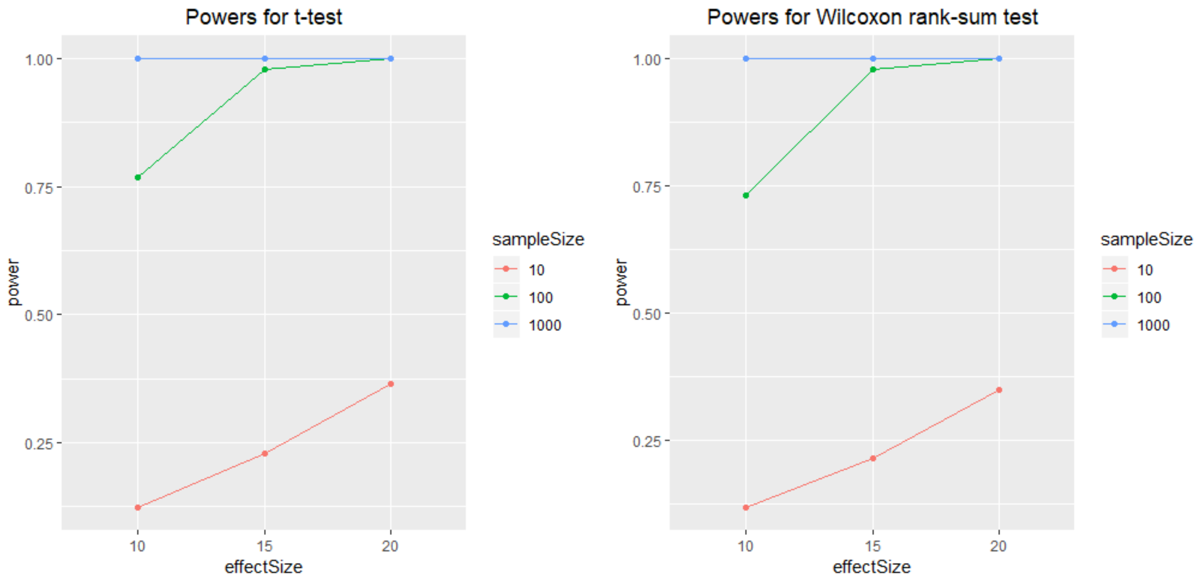

Figure3. Powers for t-test and Wilcoxon rank-sum test under different scenarios

5. Conclusions

To sum up, there are four main conclusions have been drawn in this report. First, the sample size has positive correlations with sizes as well as powers of both tests. Second, the number of decimal places has little influence on the sizes of both tests. Third, the large effect size helps to increase powers of tests. Finally, when the sample size is big enough, the powers of tests can reach to 1. However, the number of scenarios in this study is limited, the variation tendencies of the size and power might not be very convincing. This might be the direction in the further study.

Reference:

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2018). dplyr: A Grammar of Data Manipulation. R package version 0.7.7. https://CRAN.R-project.org/package=dplyr

H. Wickham. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Mahadevan, S. (1997). Monte carlo simulation. *Mechanical engineering-new york and basel-marcel dekker-*, pp.123-146.

Rafael A. Irizarry (2018). dslabs: Data Science Labs. R package version 0.5.1. https://CRAN.R-project.org/package=dslabs

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA.