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

Abstract

This study uses computer intensive statistics to investigate the size and power of a parametric and non-parametric test. A Monte Carlo simulation is used to test the power and size of a Two-Sample t-test as well as a Mann-Whitney U test for different scenarios. The parameters for those tests were inspired from the *Gapminder* dataset taken from the *dslabs* (Irizarry 2018)[1] library. The tests were chosen in accordance to the following research question: Is the life expectancy the same in rich and poor countries? The size of the parametric test is closer to the alpha value than the non parametric test, which proves that the latter is less reliable when the assumptions are met. Moreover, when the null hypothesis is violated, in the scenario where the data is not normally distributed, the Mann-Whitney U test still doesn't show a higher power value compared to the t-test for the same conditions, stating that the simulated distribution is close enough to being normal, since usually the non-parametric test is preferred in these situations.

1. Introduction

Before performing data analysis on a given dataset it's always recommended to check for the size and power of that data in order to pick the most appropriate test. Simulating Monte Carlo datasets using properties from the original data gives an overview on whether a parametric or non-parametric would be more appropriate depending on the power of the scenarios. In some cases the data would show that none of the tests can give us significant information, therefore some data transformation should be made.

This study tries to demonstrate the different power and size outputs for different scenarios, using the *Gapminder* dataset taken from the *dslabs* (Irizarry 2018)[1] library. To know whether the life expectancy is the same in rich and poor countries, the following parameters were used: life_expectancy, population, GDP.

2. Methods

Some data analysis was performed on the original data. First the GDP/Capita was calculated by dividing the GDP with the Population for every record. A summary table shows some meaningful statistical values about the GDP/Capita. The 3rd quantile was chosen as a baseline for dividing the countries into rich and poor groups. The figure 1 shows the different distributions of the data. The life expectancy histograms for each country type look slightly skewed to the left but can be considered normally distributed. Since the model has only two groups, the dependent variable (life expectancy) is continuous, the independent variable is a factor with two levels and the data is close to being normally distributed, the appropriate parametric test would be the two-sample t-test. Therefore the equivalent non-parametric test would be the Mann-Whitney U test. The normality and the homoscedasticity assumptions were then tested.

In order to evaluate the power and size of the selected tests, four scenarios were analyzed. Two of them were conducted with the purpose of examining the size, this was done by having equal means but different standard deviation and significant levels. These scenarios can give an insight of the statistical size of the tests since they do not violate any assumption. However, to analyze the statistical power, the assumptions had to be violated by simulating different means and by changing the data to a beta distribution in two other simulations.

To carry out all these computer intensive scenarios, some functions were optimized by parallelizing some computations using a cluster. The performance enhancements can be seen by using the profiling functions. The code was implemented using R studio [2].

3. Results

3.1.Scenario 1

In the first scenario, the simulated data had the same mean. The standard deviation and the sample sizes increased, the following results were obtained.

Table 1: Scenario 1 size for the parametric test

Sample Size	SD + 1	SD+1.2	SD + 2	SD + 4
10	0.052	0.057	0.047	0.051
100	0.048	0.051	0.052	0.041
300	0.053	0.056	0.053	0.055
1000	0.050	0.063	0.045	0.056

Table 2: Scenario 1 size for the non-parametric test

Sample Size	SD + 1	SD+1.2	SD + 2	SD + 4
10	0.048	0.049	0.041	0.037
100	0.049	0.047	0.053	0.043
300	0.051	0.054	0.050	0.051
1000	0.056	0.055	0.048	0.049

3.2.Scenario 2

In the second scenario, the data was simulated so that the null hypothesis would be rejected. An effect size was added to the means. the power was measured for different sample sizes as follow:

Table 3: Scenario 2 power for the parametric

Sample Size	ES + 1	ES +1.2	ES + 3	ES + 6
10	0.060	0.052	0.106	0.270
100	0.110	0.139	0.533	0.988
1000	0.608	0.786	1	1

Table 4: Scenario 2 power for the non-parametric

Sample Size	ES + 1	ES +1.2	ES + 3	ES + 6
10	0.054	0.043	0.088	0.237
100	0.109	0.127	0.512	0.988
1000	0.593	0.774	1	1

3.3.Scenario 3

In the third scenario, different significant levels were used to compare the size values.

Table 5: Scenario 3 size for the parametric

Sample Size	α =0.05	α =0.06	α =0.08	α =0.1
10	0.050	0.052	0.068	0.089
100	0.053	0.063	0.077	0.090
1000	0.049	0.065	0.085	0.097

Table 6: Scenario 3 size for the non-parametric

Sample Size	α =0.05	α =0.06	α =0.08	α =0.1
10	0.055	0.054	0.066	0.082
100	0.049	0.059	0.084	0.086
1000	0.050	0.060	0.086	0.095

3.4.Scenario 4

In the last scenario, the normality assumption was violated by using a distribution similar to the original data, by using as min, max, mean and sd the properties calculated form the original dataset. Multiple effect sizes were measure for the non normal distribution to inspect the power of the non-parametric test compared to the two sided t-test. The figure 2 shows the variation of power for the two statistical tests in relation to the effect size.

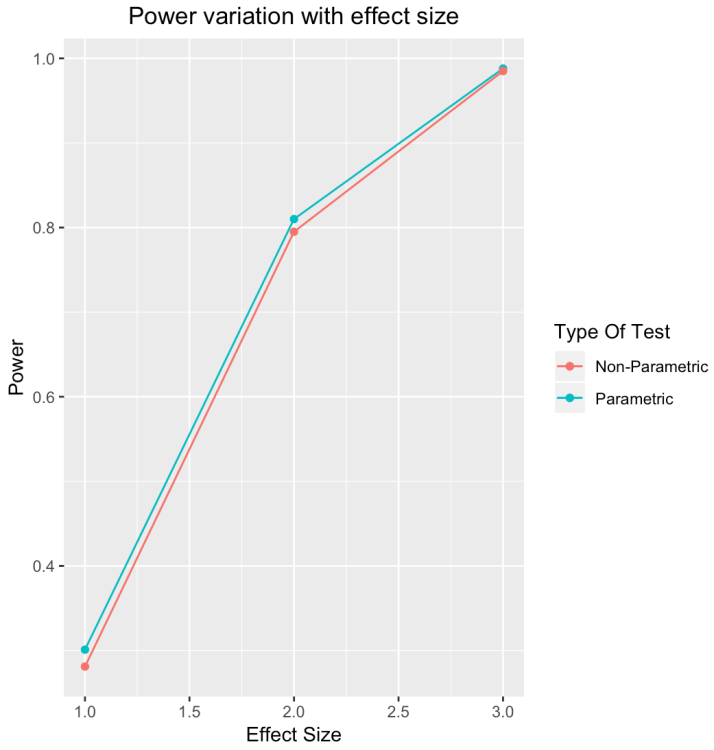

Figure 2: Power vs Effect Size for a non-normal distribution

4. Discussion

The first scenario isn't very informative, no specific pattern can be discerned. The size varies around the alpha value of 0.05 which is logical since alpha represents the probability of rejecting the null hypothesis given that it is true. In the second scenario, comparing the power for the same sample size in both tests shows that the parametric test outperforms the non-parametric one for a normal distribution. The power increases when the effect size is larger as well as when the sample size increase. This is due to the fact that larger differences in means can be detected more often, and for a fixed effect size the different means can be found in a bigger sample size.

Higher significance levels lead to rejecting the null hypothesis more often. This can be seen in the third scenario when the size increases with the value of alpha. Choosing a higher significance level can lead in falsely rejecting the null hypothesis.

Using a non normal distribution violates the two sample t-test assumptions. In this case the Mann-Whitney U test performs better than the parametric test. For the last scenario, the chosen distribution was close enough to the original data since the simulated data took into account the range of the values. The figure 2 shows that the parametric is still better than the non-parametric test which implies that the non-normality isn't strongly violated in the original observations.

5. Conclusion

The statistical size and power is an important criteria for choosing the right statistical test. For the Gapminder dataset studied in this report, the two sided t-test was shown more appropriate than the Mann-Whitney U test in all the cases.

6. References

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

[2] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Used libraries in the code:
dplyr: Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2018). dplyr: A Grammar of Data Manipulation. R package version 0.7.6. https://CRAN.R-project.org/package=dplyr

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

car: John Fox and Sanford Weisberg (2011). An {R} Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage. URL: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

MASS: Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

truncnorm: Olaf Mersmann, Heike Trautmann, Detlef Steuer and Björn Bornkamp (2018). truncnorm: Truncated Normal Distribution. R package version 1.0-8. https://CRAN.R-project.org/package=truncnorm

gridExtra: Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

foreach: Microsoft and Steve Weston (2017). foreach: Provides Foreach Looping Construct for R. R package version 1.4.4. https://CRAN.R-project.org/package=foreach

doParallel: Microsoft Corporation and Steve Weston (2018). doParallel: Foreach Parallel Adaptor for the 'parallel' Package. R package version 1.0.14. https://CRAN.R-project.org/package=doParallel

7. Appendix

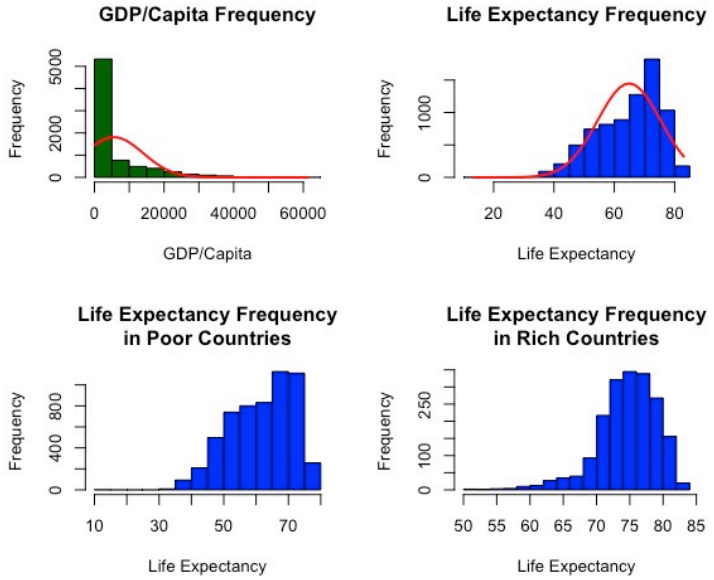

Figure 1: Histograms of the different data distributions