Statistical power and size - results from Monte Carlo simulations

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

1. Abstract

The current report aims to investigate the size and power of two statistical inferential tests - one under the parametric family of tests and one non-parametric. This analysis is enabled by utilising Monte Carlo (MC) simulations. This method allows for the investigation to simulate these statistical properties by performing numerous tests for an equal amount of different datasets while under varying scenarios. MC techniques may also be directly applied to other areas of statistical evaluation and problem-solving/estimation approaches.

2. Introduction

For this report, the initial step was to formulate a research question. From the publicly available datasets within R, the `movielense` dataset from the `dslabs` package (Irizarry, 2018) was chosen, from which such a question was formed. The current research question is:

" Are the average ratings for 21st century animated movies different from Sci-Fi movies of the same period? "

3. Methodology

a. Data characteristics

The data was extracted from the `movielense` dataset and currently contains the recorded ratings for two genres of movies - Animated and Sci-Fi movies released after the year 2000. The properties (mean and standard deviation) of these genre ratings were extracted. The ratings themselves were given on a scale of 0-5, with increments of 0,5, where 5 is a top ranking.

b. Statistical tests

"When the population is approximately normally distributed, the two-sample t-test is appropriate to conduct a hypothesis test for the difference between two means. However, when the population is not normally distributed, the two-sample t-test has low efficiency. The Wilcoxon-Mann-Whitney U two-sample test or the Kruskal-Wallis test can be considered." (Jett & Speer, 2016, p.1)

Following Jett & Speer (2016), in order to approach the current research question and to compare group means appropriately, a parametric Welch t-test and the non-parametric Mann-Whitney U test, were chosen to fulfill the criteria of comparing size and power of one parametric and one non-parametric test.

- Welch Two Sample t-test (Parametric)
    - Null hypothesis - The two means are equal
    - Alternative hypothesis - The two means are not equal
- Mann–Whitney U test (Non-parametric)
    - Null hypothesis - The distributions of both populations are equal
    - Alternative hypothesis - The distributions of both populations are not equal

c. Workflow

To develop MC simulations to achieve the sought analysis, a diagram was designed to visually represent the path by which this aim could be reached and to indicate different variations in scenarios which could be coded.

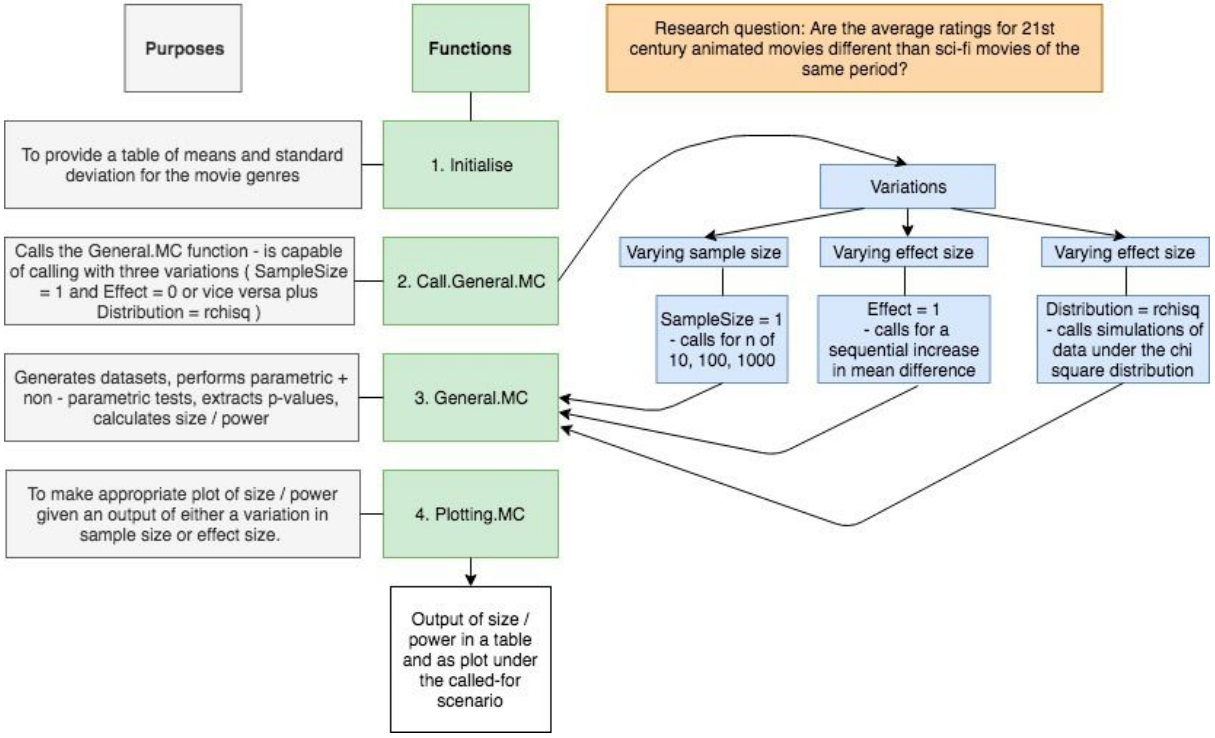


Figure 1 - Visualisation of workflow and code outline with description of functions

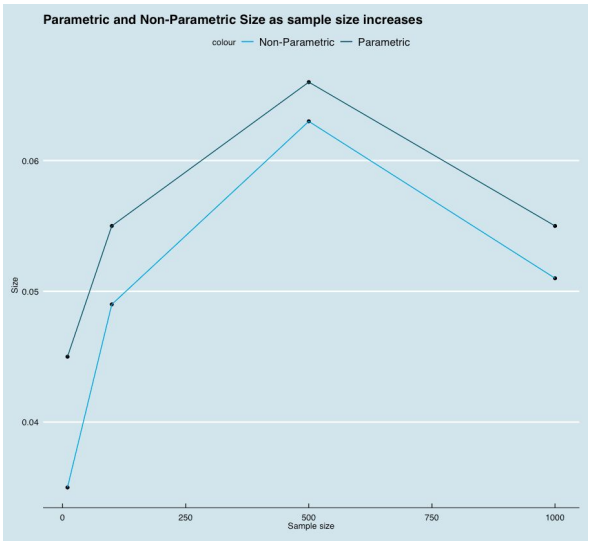
Coding one function (General.MC) to perform the MC simulations and to calculate size / power and having another function (Call.General.MC) to serve as a call-function of the first one was thought to be an appropriate way to perform the required analysis and achieve the aim. This call-function serves as the switch-table for the user to call for what output is sought - by putting either 'Effect' or 'SampleSize' as equal to 1 (instead of 0). This alters the output would be either the power of each test as (1) the sample size increases from 10 to 1000 or (2) the effect size increases (the means become sequentially more different). A final variation is to call the MC function for data generated under a non-normal distribution, e.g. a Chi square distribution. The purpose of this would be to test if the parametric test performs worse than the non-parametric when its assumption of normality is not met.

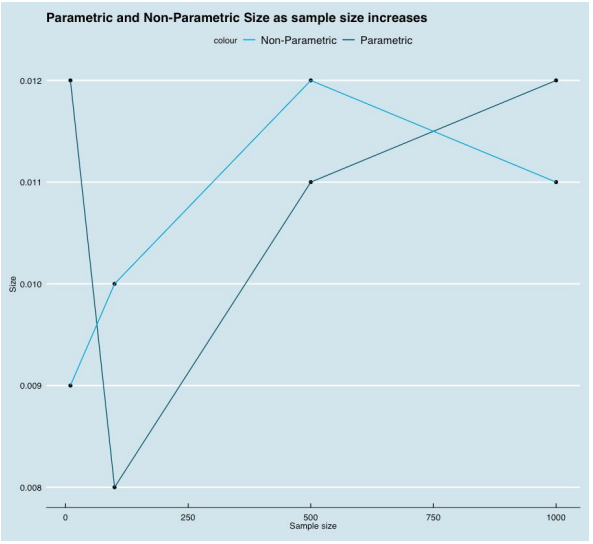
d. Scenarios

For the first two scenarios size is investigated, thus the means and standard deviation between the two samples are kept equal for the null hypothesis to be true. The first scenario simulates the size when the alpha level is 0.05, the second scenario uses an alpha of 0.01. For the third scenario the simulation is called with the same mean and standard deviation, the difference in means is then sequentially increased from 0 to 0,6 'rating-units' to demonstrate that power is influenced heavily by effect size - true differences in means between groups. The fourth scenario is the same as the third but the data is now generated

from a Chi square distribution and thus violates the assumption of normality for the Welch t-test. This then simulates how the two tests perform (power) with data from a non-normal probability distribution. The final scenario simulates the tests with the 'true' properties as obtained from the original dataset and demonstrates how power increases with sample sizes.

4. Results

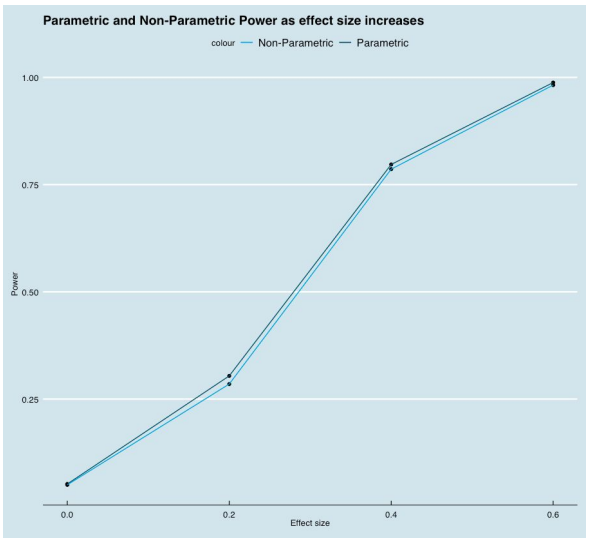


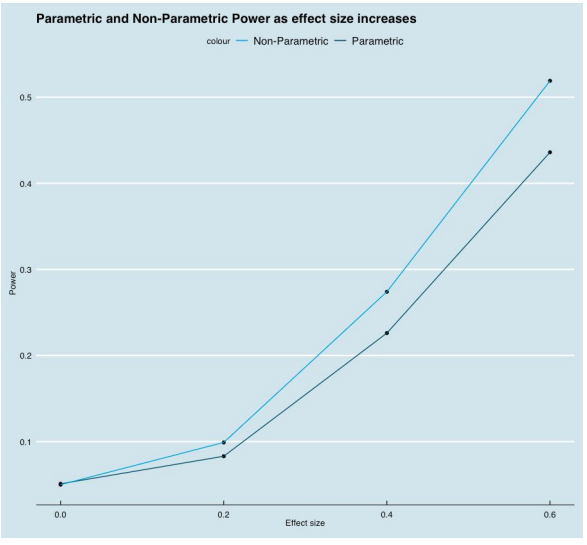


Size	Parametric	Non-Parametric
n = 10	0.045	0.035
n = 100	0.055	0.049
n = 500	0.066	0.063
n = 1000	0.055	0.051

Size	Parametric	Non-Parametric
n = 10	0.012	0.009
n = 100	0.008	0.010
n = 500	0.011	0.012
n = 1000	0.012	0.011

Figure 2 - Simulation of size for non-parametric and parametric test under the alpha level of 0.05 (**left**) and 0.01 (**right**) with corresponding tables.

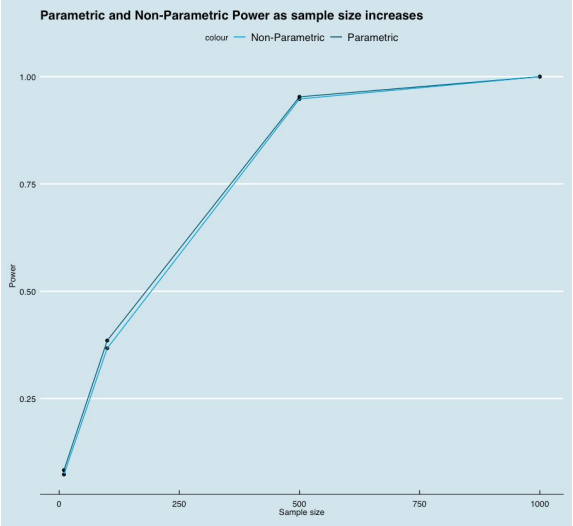




Power	Parametric	Non-Parametric
Effect = 0	0.052	0.050
Effect = 0.2	0.304	0.285
Effect = 0.4	0.797	0.786
Effect = 0.6	0.988	0.982

Power	Parametric	Non-Parametric
Effect = 0	0.051	0.050
Effect = 0.2	0.083	0.099
Effect = 0.4	0.226	0.274
Effect = 0.6	0.436	0.519

Figure 3 - Simulations of power for non-parametric and parametric test under the effect size of 0 to 0.6 'rating-units' with corresponding table where (**left**) is data simulated under a normal distribution and (**right**) under a Chi square distribution with corresponding table.



Power	Parametric	Non-Parametric
n = 10	0.083	0.073
n = 100	0.385	0.367
n = 500	0.953	0.948
n = 1000	1.000	1.000

Figure 4 - (**left)** Simulation to showcase how power increases with sample size, here the means and standard deviations are generated as under the properties of the original data / genres (Animated and Sci-Fi).

5. Conclusion

From the current Monte Carlo simulations and scenarios one can conclude that the size of both parametric and non-parametric tests tends to vary around the level of alpha (see Figure 2 left and right). The power of both parametric and non-parametric tests increases with the sample size (see Figure 4) and also with increases of effect size where the differences become incrementally different (see Figure 3 left and right).

It does appear, from these figures and tables, that for these simulations and scenarios the parametric tests tend to be associated with power above that of the non-parametric test. The exception to this is the fourth scenario (see Figure 3 right) where the normality assumptions of the parametric test were intentionally violated by simulating data under a Chi square probability distributions and not a normal distribution. In this instance the power of the parametric Mann-Whitney U test was, under simulations, greater.

A notation which would be interesting to simulate further, is how these tests relate to making Type 1 errors as given by Figure 2, these do tend to scatter around the level of alpha to some extent.

6. References

Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse

Jeffrey B. Arnold (2018). ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'. R package version 4.0.1. https://CRAN.R-project.org/package=ggthemes

Jett, D., & Speer, J. L. (2016). Comparison of parametric and nonparametric tests for differences in distribution. *2016 NCUR*.

R: A Language and Environment for Statistical Computing, R Core Team, R Foundation for Statistical Computing, Vienna, Austria (2018), https://www.R-project.org/

RStudio Team (2018). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/

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