School of Mathematics and Statistics

Computing in Statistics

A study of the size and power of parametric and non-parametric t-tests when applied to simulated datasets.

Word Count: 1939

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

Contents

Abstract

An investigation is undertaken to compute the size and power of hypotheses tests for two sample sizes which relate to the polls of presidential candidates during the 2016 US election. Initial analysis indicate that the dataset fails to meet all underlying assumption required of normally distributed dataset. Nevertheless, simulations are conducted to generate a new normally distributed dataset. Further analysis also suggests that the sample size and power are proportionally associated, meaning higher sample sizes decreases the probability of committing a Type II error. Moreover, the result reveals that size of the hypothesis test (probability of rejecting a true null hypotheses) tend to decrease as sample size increases.

1. Introduction

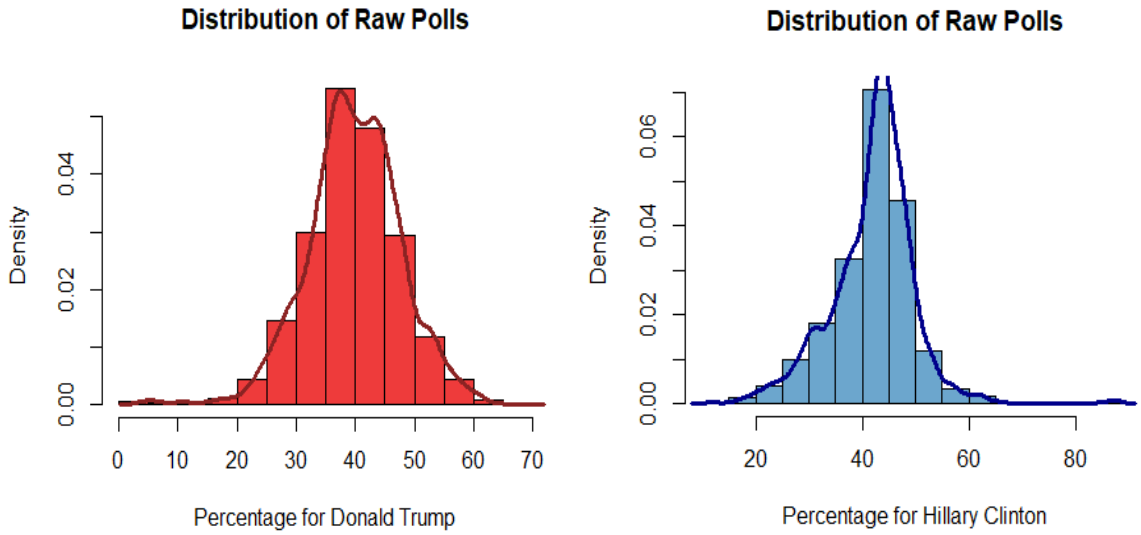
The main purpose of this study is to conduct Monte Carlo simulation which involves random sampling from probability distribution to examine the size and power of the statistical tests under different situations. The dataset used in this investigation is derived from the '*dslabs'* package in R Programming Language (Irizarry 2018). It contains the Polls results from US 2016 presidential elections which are collected from 'HuffPost' Pollster, 'RealClearPolitics' and various polling firms. This study begins by diagnosing the assumption required for normally distributed dataset. It then moves on to formulate a two-sided hypotheses test, and performs statistical t-tests which are later applied to the simulated data. In the end, it draws comparison between the statistical power and size of the tests for both normal and non-normal dataset. All exploratory analysis are carried out with the help of R Programming Language using RStudio IDE software application.

2. Methodology: An initial observation of dataset provided (polls_us-election_2016) indicates there are n = 4208 observations in total with 15 columns. Full definitions for each column can be found in Appendix I. In this study, six columns are identified with no new or supportive information for the analysis. The columns regarding raw polls of candidates such as Gary Johnson and Evan McMullin are eliminated from dataset as they contains no predictive value. This study, therefore, only uses the percentage of opinions polls of Donald Trump and Hilary Clinton to perform the statistical t-tests. Before using the dataset to carry out statistical t-tests, the following assumptions are examined:

Test for normality of sample sizes is performed using the Quantile-Quantile (QQ) plot (Ho: the data for each group appears to have come from a Normal distribution). The plots illustrate that two groups roughly follow a straight line (see Appendix II). Histograms (Figure 1) are used to visualise this result and show the distribution of 'raw polls' for both Donald Trump and Hilary Clinton in the dataset. Figure 1 generally indicates that two groups follow a normal distribution. Tests for constant spread are carried out. The Null Hypothesis (Ho) is that the underlying standard deviations for each group appear to be equal. Since the p-value result is less than 0.05 the null hypothesis can be rejected, meaning the constant spread assumption doesn't hold true (see Appendix II). Tests for independence are completed using a chi-square test. The null hypothesis is that knowing the level of Variable x does not help to predict the level of Variable y. That is to suggest, the variables are independent. Since the p-value result is less than 0.05 the null

hypothesis can be rejected. Therefore, assumption of independence for the two groups does not hold true (see Appendix II).

Figure 1. Histograms of an approximately normally distributed variables



2.1 Research Question

In this study two applications of the t-test are performed. These tests are comparing the means of two groups in dataset namely the percentages of Donald Trump and Hilary Clinton in the opinion polls throughout the 2016 presidential election across the United States. The hypotheses for this question are formulated as follows:

H0: $\mu 1 = \mu 2$, $\mu 1 - \mu 2 = 0$ or in plainer language: the means of the two groups that have been sampled are equal. That is, the samples are just drawn from the same population.

H1: $\mu 1 \neq \mu 2$, $\mu 1 - \mu 2 \neq 0$ or in plainer language: the means of the two groups that have been sampled are not equal, suggesting the samples are drawn from statistically distinct populations.

To test the null hypothesis, both parametric and non-parametric t-tests are conducted where x1 and x2 are the sample means and s is the standard error of the difference between the means. A parametric statistical test assumes that the corresponding data population distributions follow the normal distribution, while a non-parametric test does not make such assumptions. That said, the statistical tests often don't perform as expected and might incorrectly reject the null hypothesis. This is particularly likely to occur if some of the underlying assumptions are not satisfied. The error of rejecting a null hypothesis when it actually holds true is called Type I error, also known as a "false positive". The size of a test calculate the probability of committing a Type I error when H0 is actually true. On the other hand, incorrectly failing to reject a null hypothesis

when the alternative hypothesis is the true state of nature is called Type II error or "false negative". The power of any statistical test thus measures the probability of not making a Type II error.

3. Results

In this case study, the paired sample *t*-test (parametric) is used to determine whether the mean difference between two groups is zero since the dataset only appear to be approximately normally distributed. Since the p-value from t-test is less than 0.05 the null hypothesis can therefore be rejected. That is to say, mean difference of both groups dose not equal zero. Findings are detailed in Table A. In addition to this procedure, the Wilcoxon Signed-Rank Test (non-parametric) is performed to test the null hypothesis as the dataset fails to meet constant spread and independence assumptions. The result reveals that mean differences of both group is not equal to zero, and therefore the alternative hypothesis holds true (see Table A).

Table A. Results of Statistical Tests

Table 1. Two Sample t- Test (parametric)

t-Test	Degree of Freedom	P-Value
12.709	8414	2.2e-16

Table 2. Wilcoxon Signed-Rank Test (non-parametric)

Data	V	P-Value
Polls of 2016 US Election	5306000	2.2e-16

To devise a computer intensive approach this investigation then simulate samples drawn from the dataset. Simulations generates random sample sizes which follow a normal distribution, based on a given mean and a standard deviation. It then analyses each group and compute the proportion of results that are significant for both parametric and non-parametric t-tests. That proportion is the estimated power for the hypotheses test. If the computed p-value is equal to or smaller than some pre-defined value, then it is unlikely that the data could be generated under H0. The null hypothesis therefore is rejected. The pre-defined value (α-level) by convention is set to equal α =0.05. In a similar vein, the size of the hypothesis test is computed for p-values greater than 0.05 (α-level), for any statistical t-tests. The findings of simulation process are reported in Table b. As *the sample size* increases the probability of not committing a Type II error increase (see Figure 2). That is to say, the hypothesis test performs well in identifying a false null hypothesis. Moreover, the parametric test (two sample t-test) tends to have higher power compared with non-parametric tests, when the dataset is normally distributed.

Other factors such as variance, effect size and alpha level tend to influence the proportion of statistical power. For example, effect size shows how relevant is the relationship between two sample sizes provided in the dataset. In this case study, the effect size is measured based on differences in averages of each sample size which is known as Cohen's D (the rule of thumb states that small effect = 0.2, Medium Effect = 0.5, Large Effect = 0.8). When the Cohen's D increases, the proportion of power also does increase as shown in Figure 2. Furthermore, there is a positive relationship between the significance level (alpha) and power. As regards the size of hypothesis test the statistical result confirms that there is an inverse relationship between Type I and Type II errors: as one increases the other decreases. Thus as sample size increases the probability of rejecting a true null hypothesis tends to decrease.

Table b. Results of Estimated Power and Size from normally distributed dataset

Table 1. Power estimated from Wilcoxon t-test

Sample Size	Effect Size	Power
100	0.2	0.467
100	0.5	0.491
100	0.8	0.429
250	0.2	0.844
250	0.5	0.846
250	0.8	0.850
750	0.2	0.998
750	0.5	0.999
750	0.8	1.000

Table 2. Power estimated from two sample t-test

Sample Size	Effect Size	Power
100	0.2	0.494
100	0.5	0.482
100	0.8	0.490
250	0.2	0.871
250	0.5	0.859
250	0.8	0.873
750	0.2	0.999
750	0.5	1.000
750	0.8	1.000

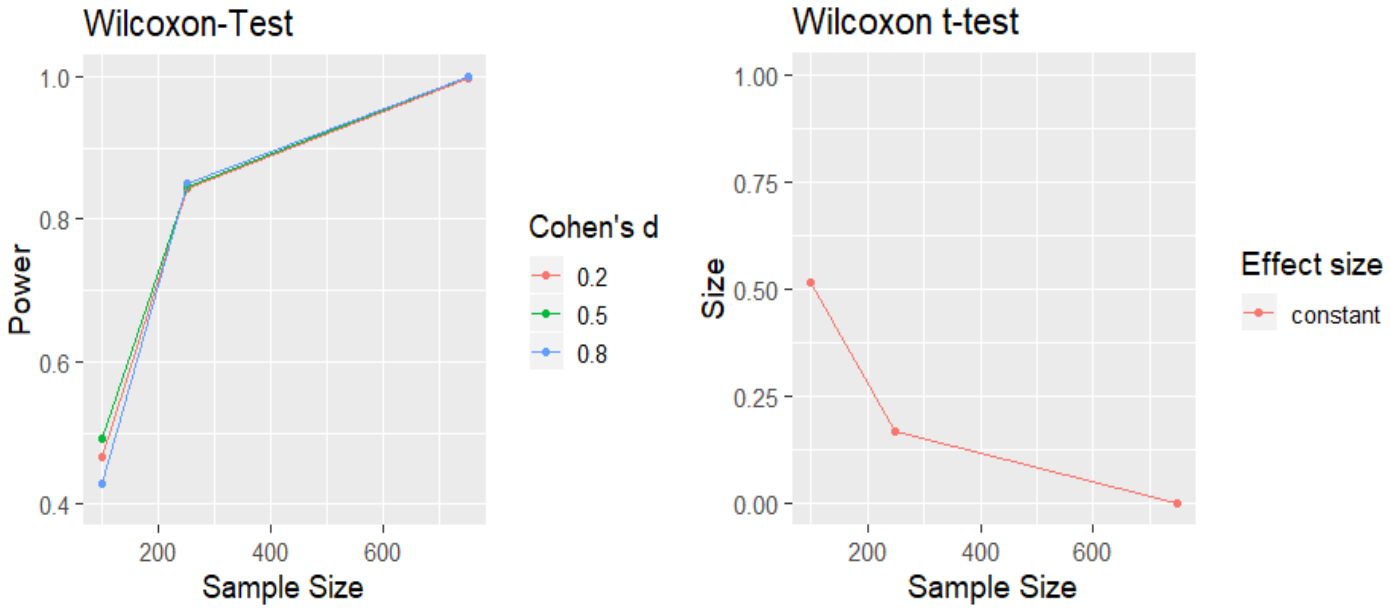
Table 3. Size estimated from two sample t-test

Sample Size	Effect Size	Power
100	Constant	0.505
250	Constant	0.133
750	Constant	0.001

Table 4. Size estimated from Wilcoxon t-test

Sample Size	Effect Size	Power
100	Constant	0.517
250	Constant	0.170
750	Constant	0.000

Figure 2. Estimated Power and Size of Wilcoxon t-test



Further investigation reveals the statistical power of a randomly generated dataset that follows chi-square distribution tends to be relatively lower compared with a normally distributed dataset. The findings are detailed in Table c. From this analysis, it can be argued that statistical power computed from a non-parametric t-test which follows normally distributed sample sizes tend to have more statistical power than those non-normally distributed. That is say, one is more likely to detect a significant effect when the dataset follows a normal distribution.

Table 1. Comparison between Normal and non-Normal distribution

Table 5. Power estimated based on non-normal dataset

Sample Size	Effect Size	Power
100	0.2	0.298
100	0.5	0.265
100	0.8	0.318
250	0.2	0.589
250	0.5	0.604
250	0.8	0.582
750	0.2	0.975
750	0.5	0.980
750	0.8	0.982

Table 1. Power estimated from Wilcoxon t-test

Sample Size	Effect Size	Power
100	0.2	0.467
100	0.5	0.491
100	0.8	0.429
250	0.2	0.844
250	0.5	0.846
250	0.8	0.850
750	0.2	0.998
750	0.5	0.999
750	0.8	1.000

4. Conclusion

The main goal of this investigation was to generate a random dataset that follows a normal and non-normal distribution with the purpose of computing power and size of the test hypothesis (Whether mean difference of two polls equal zero) under a number of scenarios. Preliminary analysis revealed that two sample t-test offers higher statistical power than Wilcoxon t-test. One explanation is that parametric t-test assumes that data follows a normal distribution. Moreover, a proportional relationship between sample size, significance level and statistical power were identified. Similarly, the probability of committing Type I error declined as number of sample size increased in the simulation. It's commonly assumed that a need to select between a parametric and nonparametric test happens when dataset follow or does not normal distribution. However, additional exploration indicated that a normally generated dataset presents higher statistical power for non-parametric t-test than those that follow a chi-square distribution for instance.

5. References

Fabozzi, F.J., Focardi, S.M., Rachev, S.T. and Arshanapalli, B.G., (2014). *The basics of financial econometrics: Tools, concepts, and asset management applications*. John Wiley & Sons.

James, G., Witten, D., Hastie, T. and Tibshirani, R., (2013). *An introduction to statistical learning* (Vol. 112). New York: springer.

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

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

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

Appendix I - Variables in data file

- state: State in which poll was taken. `U.S` is for national polls.
- startdate: Poll's start date.
- enddate: Poll's end date.
- pollster: Pollster conducting the poll.
- grade: Grade assigned by fivethirtyeight to pollster.
- samplesize: Sample size.
- population: Type of population being polled.
- rawpoll_clinton: Percentage for Hillary Clinton.
- rawpoll_trump: Percentage for Donald Trump
- rawpoll_johnson: Percentage for Gary Johnson
- rawpoll_mcmullin: Percentage for Evan McMullin.
- adjpoll_clinton: Fivethirtyeight adjusted percentage for Hillary Clinton.
- ajdpoll_trump: Fivethirtyeight adjusted percentage for Donald Trump
- adjpoll_johnson: Fivethirtyeight adjusted percentage for Gary Johnson
- adjpoll_mcmullin: Fivethirtyeight adjusted percentage for Evan McMullin.

Appendix II - Figures and Tables from checking assumptions

Table 1. Chi-square Independence Test

X-squared	Degree of Freedom	P-Value
3868200	1814400	2.2e-16

Figure 1. Q-Q Plots for Trump polls and Hillary Polls