Assignment 3 Report

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

Abstract

In order to determine the size and power of two statistical tests, the randomization test and the two sample t-test, data was simulated that is similar to the brexit_polls dataset found in the R package dslabs was simulated and the tests were conducted on the spread (proportion of remain – proportion of leave voters) data simulated with the null hypothesis of the tests being that the mean spread was the same for polls that were conducted online and those that were conducted on the telephone. The data was simulated by modelling the number of remain and leave voters as multinomial with different probabilities depending on whether a poll was conducted online or by telephone and drawing random deviates and then calculating the proportions of each type of vote (Remain, leave and undecided) and subtracting the remain proportion from the leave proportion to get the spread for each. The data was simulated for cases in which this null hypothesis was true and false for 3 different scenarios of data collection, one that looked at what effect: the number of polls conducted had, one that looked at what effect the sample sizes of the polls had and one that looked at what effect the difference in spreads had on the power and size of the tests. It was found that the power of the test was 1 for all scenarios for both tests meaning that when the null hypothesis was false both tests always correctly rejected the null hypothesis. The randomization test was found to have small sizes for every scenario which means that it did not incorrectly reject the null hypothesis very often. However, it was found that the t-test had large sizes for each scenario of data collection. This showed that the two-sample t-test often rejects the null hypothesis incorrectly meaning that it is not a particularly good method to use for analysing data of this form. It was found that the t-test had a smaller size when the sample size is generally lower in the polls but performs that differing the number of studies or the size of the difference in spread between the two groups had no effect on the size of the two-sample t-test. It was found that for the randomization test that having generally larger sample sizes for the polls, having more polls conducted and having a larger difference in probabilities increased the size of the randomization test. Overall, the randomization test should be preferred for testing if the spread means are the same and it gives more accurate results if there are less polls conducted, the number of participants in the polls is smaller and the difference in spreads between the polls is smaller.

Introduction

The aim of this project was to test the power and size of statistical tests by testing them on simulated data. The dataset I have chosen to mimic with my simulated data is the brexit_polls dataset from the R package dslabs which gives the results of various online and telephone polls about whether people are voting remain, leave or are undecided when it comes to voting in Brexit. The research question I have chosen to examine is: Does the method of data collection (Online or telephone) affect the difference between leave and remain votes for Brexit data? I will be testing the power and size of the Two Sample T-Test and the randomization test with the null hypothesis of

both being that the mean spread (proportion of remain voters – proportion of leave voters) for both type of polls is the same.

Method

Simulating the Data

A summary of the original data found in the brexit_polls dataset is shown in the table below:

Column Name	Represents	Data Type	Range
startdate	Date Poll Started	Date	08/01/2016 - 23/06/2016
enddate	Date Poll Ended	Date	10/01/2016 - 23/06/2016
pollster	Company that conducted poll	String	N/A
poll_type	Whether poll was conducted online or by telephone	String	N/A
samplesize	Sample size of poll	Integer	497 – 4772
remain	Proportion of sample who voted remain	Floating Point	0.35 – 0.55
leave	Proportion of sample who voted leave	Floating Point	0.32 – 0.55
undecided	Proportion of sample who were undecided	Floating Point	0 – 0.3
spread	remain - leave	Floating Point	-0.1 – 0.19

In order to simulate data similar to this data, I decided to model the counts of remain, leave and undecided voters as multinomial data such that $(\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}) \sim Multinomial(p_1, p_2, (1 - p_1 - p_2))$ with $\boldsymbol{x_1}$ being the count of remain voters for a poll, $\boldsymbol{x_2}$ being the count of leave voters for a poll and $\boldsymbol{x_3}$ being the count of undecided voters. I created a simulation function that took the number of studies that are online and telephone, the sample size of these studies, the probabilities of a voter being remain or leave (given by p_1 and p_2 respectively) for an online poll and the probabilities of a voter being remain or leave for a telephone poll. The function randomly generates samples from a multinomial distribution for each poll with the inputted probabilities given that the poll is either conducted online or by telephone. The function then finds the proportion of these samples that are from being remain, leave or undecided and calculates the spread, which is the proportion of samples that voted remain minus the proportion of samples that voted leave. It then outputs a data frame where each row gives the type of poll (Online or Telephone), the sample size of each poll and the spread for a one simulated poll. The design of the simulation process is shown in Appendix and a histogram of the spread of a sample simulated using my function can be seen in Figure 2. In this general simulation p_1 and p_2 are 0.4 and 0.3 for both types of poll, there are 110 polls, 55 of each type and the sample sizes of each study were simulated from a normal distribution with mean 400 and a variance of 100.

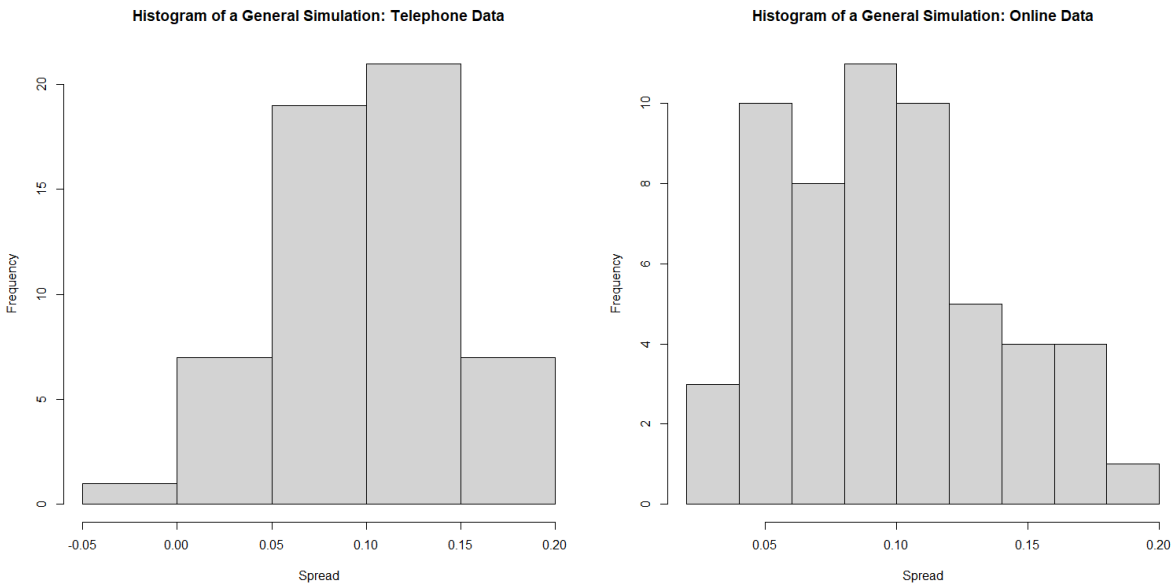

Figure 2

Testing Statistical Tests

I will be using the tests on the spread data to identify if there is any difference in spread between the online and telephone data. I will be testing my chosen tests for 3 circumstances of data collection. The first circumstance tests the power and size of the test given there is either a large or small number of participants in the studies. I have achieved this by simulating the sample sizes of the studies from normal distributions with large and small means. The second circumstance tests the power and size of the test given there is either a large or small number of participants in the studies. I have achieved this by inputting values much higher and much smaller than the number of studies in the original dataset. I chose to simulate 20 online polls and 20 telephone polls for the smaller amount of polls simulation and simulated 70 online polls and 70 telephone polls for the larger amount of polls simulation. The third circumstance is data where there is a large difference and a small difference in spread between the two types of study. I chose to use a spread of 0.7 for the telephone polls and -0.7 for the online polls for the situation where there is a large difference in spreads and a spread of 0.15 for telephone polls and -0.15 for online polls for the situation where there is a small difference in spreads. I will test the tests by repeatedly simulating data where the mean spread is the same for both online and telephone data and simulating data where the mean spread is the different between online and telephone data and then calculating the size and power of the tests.

Results

A table summarising the power and sizes of the tests under different scenarios of data collection is seen below in Figure 3 and the sizes of the tests for different scenarios can be seen visualised in a bar chart in Figure 4:

Scenario	Test	Size of Test	Power of Test
Small Difference in Spreads	Randomization	0.0281690	1
	T-Test	0.8732394	1
Large Difference in Spreads	Randomization	0.0422535	1
	T-Test	0.8732394	1
Small Sample Sizes For Polls	Randomization	0.0281690	1
	T-Test	0.6478873	1
Large Sample Sizes For Polls	Randomization	0.0563380	1
	T-Test	0.9859155	1
Small Number of Polls	Randomization	0.0281690	1
	T-Test	0.9577465	1
Large Number of Polls	Randomization	0.0704225	1
	T-Test	0.9577465	1

Figure 3

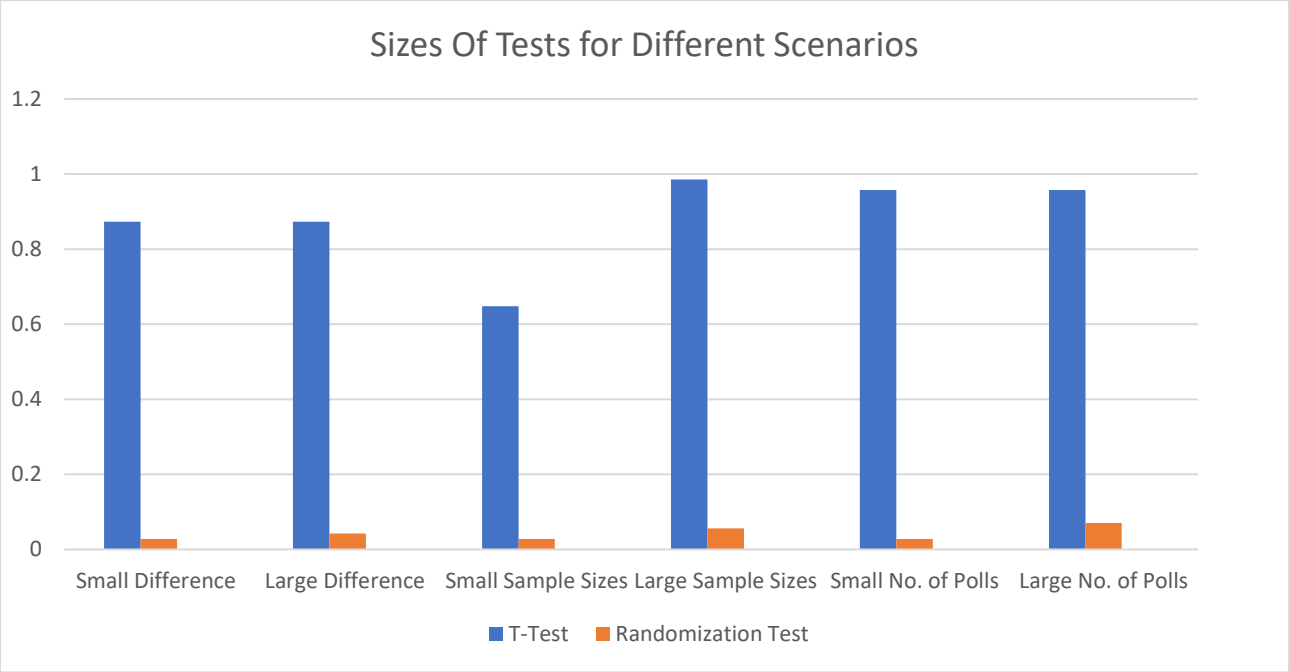

Figure 4

Conclusions

It can be seen from Figure 3 that the power of the tests is always 1 for both tests in all circumstances which means that the hypothesis that the mean spreads are the same is always accepted in the circumstance where the mean spreads are the same. Looking at the scenarios where there is either a small difference in spreads or a large difference in spreads the size of the randomization test is larger for a larger difference in spread and the size for the t-test is the same for both scenarios. This implies that the randomization test is more accurate for smaller differences in spread as opposed to larger differences in spread and there is no difference in accuracy between the two scenarios for the t-test. Next looking at the scenarios where there is either a small or large number of polls the size of

the randomization test is larger for a larger number of polls than a smaller number of polls and the size for the t-test is the smaller when there are less polls. This implies that both the randomization and two sample t-test are more accurate when there are less polls. Next looking at the scenarios where there is either small or large sample sizes for each poll the size of the randomization test is larger for a larger sample size for polls and the size for the t-test is the same for both scenarios. This implies that the randomization test is more accurate when there are smaller sample sizes for each poll as opposed to larger differences in spread and there is no difference in accuracy between the two scenarios for the t-test.

The t-test has a very large size for all scenarios with its size always being more than 0.5 meaning that it incorrectly rejects the hypothesis that online and telephone data have the same spread more than half the time. This implies that the test is probably not very well suited to this type of data. This may be because the assumption that the spread data is normal distributed for both groups may not hold which can possibly be seen in Figure 2. The t-test also assumes that the data have the same variances which also does not seem to hold for a general simulation as shown in Figure 2. The randomization test, unlike the t-test, has a generally low size and so rejects the null hypothesis a lot less when it is true. Overall, this means that the randomization test is generally more suited to analysing this question for this dataset and should be preferred when answering the same question for other similar datasets and that this test gives more accurate results when the number of polls conducted, the sample sizes of the polls and the difference in spreads are smaller.

References

Rafael A. Irizarry and Amy Gill (2021). dslabs: Data Science Labs. R package version 0.7.4. https://CRAN.R-project.org/package=dslabs

Appendix

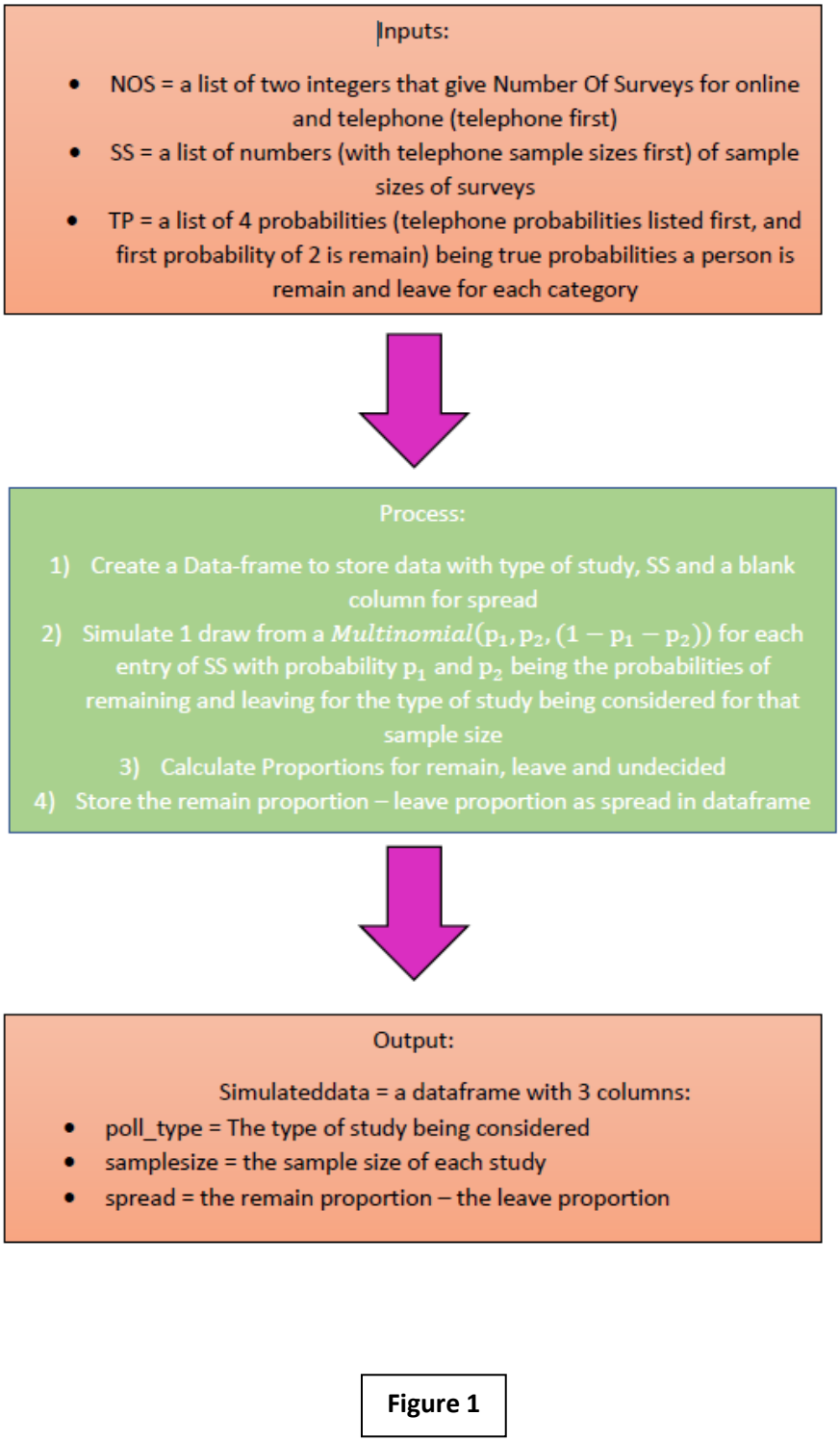

Figure 1