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

## Abstract

This report outlines a simulation study run on the “stars” data set that aims to explore whether a relationship exists between the magnitude and the temperature of a star. Data are simulated according to the properties of the observed values for temperature and following a linear regression model for magnitude to produce data that can help answer the question of whether a relationship exists between the two properties. The simulation scenarios that are explored consider different effect sizes between magnitude and temperature, different sample sizes of the data sets, and measurement errors in recording the values. Parametric t-tests and a non-parametric randomisation test is run on the simulated data sets, and from these, it is found that the ability to correctly identify the presence or absence of a relationship weakens with lower effect and sample sizes than the effect size and sample size in the observed data.

## Introduction

This report aims to simulate data within the context of an observed data set before applying statistical tests and evaluating their results. The “stars” data set from the “dslabs” package is considered, which contains the physical properties of select stars. In total, 96 observations are available that each record the name, luminosity, temperature, and spectral class of a star.

To inform the simulation study, the research question, “is there a relationship between the temperature of a star and its luminosity?” is formulated. To evaluate whether a relationship exists, simple linear regression is used, taking the temperature of a star, ‘temp’, as the predictive variable, and the ‘magnitude’ of the star, which captures luminosity, as the response. The statistical tests to measure this relationship are a t-test of the regression coefficient for ‘temp’, and a randomisation test of the coefficient from a simulated data set against a distribution generated under the assumption of no relationship. Data will be simulated for both ‘temp’ and ‘magnitude’ under varying scenarios, and the statistical tests applied. The correctness of the results of these statistical tests against what is assumed to be true is then calculated for the data simulated under each scenario. These measures of correctness are the statistical size, for when it is assumed that no relationship exists, and power, when a relationship is assumed to exist.

## Methods

To simulate appropriate values, the properties of the observed data and the relationship between the response and predictor variables must be captured. The observed distribution of the predictor variable, temperature, appears to be highly positively skewed with a small peak in the tail, as seen in Figure 1. This is replicated by simulating using a mixture model approach, splitting the distribution into two groups. The first group contains the heavy concentration of lower observed temperatures and uses a truncated normal distribution to capture what resembles the right-hand side of a bell curve shape present in the lower observed temperatures. A larger proportion of values are simulated from this group, proportional to the higher concentration of lower temperatures in the observed set. The second group of temperatures, containing a lower proportion of values that are more spread out, are captured with a normal distribution using the mean and standard deviation of values within that subgroup. This lower proportion of values that are more spread has the effect of flattening out the normal distribution in relation to the first subgroup. This results in an example set of simulated temperature values in Figure 1, which appears to capture the properties of the observed temperatures.

Chart, histogram

Description automatically generatedFigure 1: (Left) Distribution of temperature values from observed data set. (Right) Example distribution of temperature values from simulated data using observed effect size

The values for the response variable, magnitude, are then simulated from a normal distribution with a mean determined by a simple linear regression between magnitude and temperature. Hence, the mean is dependent on the size of the relationship between variables and the value of the predictor, temperature. This captures the observed distribution of magnitude according to its relationship with observed temperatures, but in a way that allows the response variable to be simulated under different scenarios and the results subsequently tested.

### 2.1 Simulation Scenarios

The simulation scenarios to be considered are shown in Table 1. First, varying effect sizes of temperature on magnitude are simulated to evaluate the effect sizes for which a significant relationship between the variables holds. Effect sizes either side of the one observed in the original data set are considered to see the extent to which a significant relationship is found. For each effect size, different sample size scenarios are considered by simulating varying sized data sets, from as small as 10 up to 500 in a data set. This allows the statistical significance of tests to be compared across a range of sample sizes to determine how samples are needed to achieve satisfactory results. This represents realistic studies whereby more or less resources allow for larger or smaller data sets, owing to the expense of the accurate study of stars. As such, in measuring the surface temperature and luminosity of stars, a scenario in which measurement error is present is considered to see if a relationship between the two holds even with some error in measurement.

Table 1: Simulation Scenarios

### 2.2 Simulation Design

To conduct the parametric t-test, testing for a significant relationship between variables, the observed data set is specified along with the relationship to test and the scenarios to test under. This is passed to a function that simulates data sets with a relationship between variables that has the properties of the simulation scenarios provided. A simple linear regression is then fit to each simulated data set, and the *p*-values for the relationship between the predictor and the response is computed and extracted. Under scenarios where the effect size is simulated as zero, the proportion of times that a statistically significant relationship was found, even though one was not assumed during the simulation, is calculated, giving the *size*. In scenarios where the effect size is nonzero, and therefore the idea of no relationship is assumed to be rejected, the proportion of times in which it was correctly rejected is calculated, giving the *power*.

For the non-parametric randomisation test, a data set is simulated under each scenario and a model fit to this data. The coefficient estimate for the relationship between variables is recorded and taken to be the observed effect size. The response values in this data set are then randomised to produce new, random data sets, and models are fit to each. Taken together, these coefficients from the models fit to randomised data give the distribution of estimates when assuming no relationship exists. The proportion of these estimates more extreme than the one observed in the initial simulated model is then calculated, giving the *p*-value for the test.

## Results and Discussion

The power of the parametric test results for the scenarios in which nonzero effect sizes were used to simulate data sets are shown in Table 2 for the different sample sizes. This shows that as the effect size between temperature and magnitude weakens, the proportion of times in which the hypothesis of no relationship is correctly rejected decreases. Conversely, for the same effect size, increasing the sample size generally increased the proportion that a relationship between the two variables was correctly identified. This indicates that collecting and simulating more data leads to a relationship being correctly detected when one is present a higher proportion of the time.

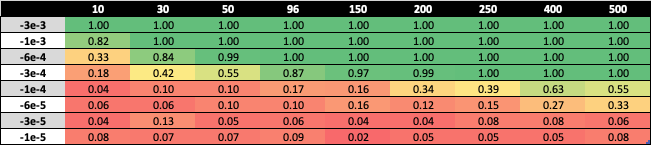
The effect size from the original model fit to the observed stars data was -6e-4. Simulating using this value and the observed sample size of 96 returns a power of 1. This is consistent with the extremely low observed probability of < 4.4e-12 of finding a relationship more extreme than the one found when assuming no relationship exists, hereby suggesting a significant relationship is extremely likely to be found with these properties for the data. Holding this observed effect size, a power of 0.8 is achieved simulating with a sample size of 25, suggesting that the observed relationship between magnitude and temperature is strong enough that reduced sample sizes can be collected whilst retaining an 80% probability of correctly rejecting the null hypothesis at the 5% significance level.

Table 2: Power results for t-tests run on simulated data with varying effect, sample sizes

Simulating data for which no effect was used between magnitude and temperature gives size results shown in Table 3. These indicate that the models fit to the data incorrectly rejected the hypothesis of no relationship around 5% of the time across all sample sizes, which is consistent with the 5% significance level used to reject or fail to reject the presence of a relationship.

The results for the parametric tests for the varying effect and sample sizes are shown in Figure 2 for nonzero and zero effects, respectively. These reflect how the power of the t-tests generally increase with larger sample sizes, but decrease for smaller effect sizes, and how the size remains around the level of significance for all sample size scenarios.



Table 3: Size results for t-tests run on simulated data with zero effect size

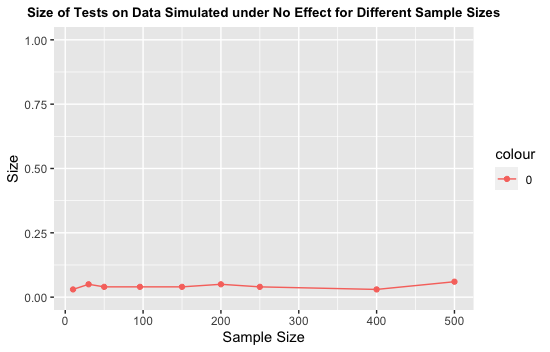
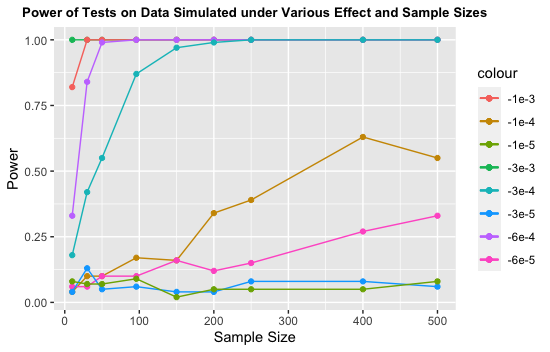


Figure 2: (Left) Power results for t-tests on simulated data for varying nonzero effect sizes. (Right) Size results for t-tests on simulated data with zero effect size.

**A screenshot of a computer

Description automatically generated with low confidence**For the scenarios with varying measurement errors, the results when rounding magnitudes to one significant figure and to the nearest five are almost identical to when no measurement error is present. This reflects the strength of the relationship between magnitude and temperature. The power results change marginally when rounding to the nearest 10, however, with differences in results highlighted in red in Table 4. The general pattern to observe is that power was reduced where a change existed, indicating that significant measurement error can lead to a lower proportion of correct conclusions of a relationship being present.

Table 4: Power results for t-tests run on simulated data with measurement error rounding to the nearest 10.

For the non-parametric randomisation test, the power results are shown in Table 5. Given that these simulations were generated with nonzero effect sizes, the coefficient estimate from the original simulated data set is taken to be the test statistic. The coefficient estimates from the repeated randomised data sets then provide the distribution to compare the original coefficient to. As seen in the results, the overall pattern of lesser effect sizes yielding lower power values holds true as it did for the parametric t-tests. It is also the case that increasing the sample size gives generally higher power values.

A picture containing text, scoreboard

Description automatically generatedTable 5: Power results for randomisation tests run on simulated data.

Table 6: Size results for randomisation tests run on simulated data.

Simulating with a zero-effect size present gives size results as shown in Table 6. For sample sizes 10, 30, 50, 150, 400, and 500, a size of 0 was observed, indicating that hypothesis of no relationship was not incorrectly rejected. There were high size values for sample sizes 96, 200, and 250, showing that there were randomised coefficient estimates that were greater than the one from the data set used for randomisation.

For data simulated with a measurement error rounding to 10, power results vary more and are generally weaker for the randomisation tests, but a similar pattern is observed, as seen in Figure 3. This suggests that whilst measurement error weakens the ability to correctly identify a relationship when one is present, collecting a large enough sample size should produce powerful results. As was the case with no measurement error, the size results are generally at 0, however outlier results exist for the sample sizes 30, 96, 200, and 250, implying issues in generating acceptable results when measurement error is present.

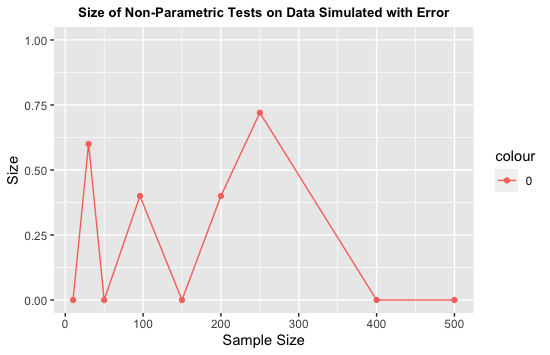


Figure 3: (Left) Power results for randomisation test on simulated data with measurement error. (Right) Size results for randomisation test on simulated data with measurement error.

## Conclusions

Overall, the patterns in the results from the simulation scenarios suggest that the simulations were run accurately. Evidence of this is the strong power and size results observed for the observed effect and sample sizes upwards and generally weaker results when the effect sizes and sample sizes decreased below the observed values. The assumptions of the observed model held true, with residuals being normally distributed, star properties being independent, and linearity between magnitude and temperature. Therefore, the parametric t-tests results have some additional validity as data were simulated from this model function. However, the non-parametric results being consistent with these results offers further support of the overall relationship between magnitude and temperature.