

STA130H1 Capstone Project Proposal

Xuanqi Wei, Shujun Yang, Riyad Valiyev, Nicolas Dias Martins

March 9, 2023

Contents

1	Introduction	0
2	Research Questions	1
2.1	RQ1: How is the galaxy's total size related to the percentage of light within the radius?	1
2.1.1	Data	1
2.1.2	Method	1
2.1.3	Visualization	1
2.2	RQ2: Do the galaxies farthest to the Earth differ significantly from the closest ones in terms of their total luminosity?	2
2.2.1	Data	2
2.2.2	Method	2
2.2.3	Visualization	2
2.3	RQ3: How well can a simple linear regression model predict the galaxy's apparent brightness from redshift?	3
2.3.1	Data	3
2.3.2	Method	3
2.3.3	Visualization	3
3	Contribution	4
4	Timeline	4

1 Introduction

The following are sections of 'Capstone Project Proposal' for STA130H1 in 2023 winter session written by Xuanqi Wei, Shujun Yang, Riyad Valiyev and Nicolas Dias Martins. It is a pleasure to sincerely acknowledge our indebtedness to our TA, whom we warmly thank for his observation and suggestions, which materially improved the proposal draft.

2 Research Questions

2.1 RQ1: How is the galaxy's total size related to the percentage of light within the radius?

2.1.1 Data

The data we will use in research question 1 comes from 'Galaxy Zoo Tabular Data Contents', more specifically, `petro_theta`, which is an estimate of a galaxy's total size in radius, `petro_th50`, which is an estimate of the galaxy's size in radius where 50% of the light is inside the radius and 50% of the light is outside the radius, and `petro_th90`, which is an estimate of the galaxy's size in radius where 90% of the light is inside the radius and 10% of the light is outside the radius. According to the aforementioned descriptions, the 'galaxy's total size' is related to 'petro_theta', while the percentage of light within the radius' is related to both 'petro_th50' and 'petro_th90'.

2.1.2 Method

The method we will use in research question 1 is `bootstrapping`. In research question 1, we will use bootstrapping to simulate sampling distributions of the mean of three sample statistics (by using the same random sample): 'petro_theta', 'petro_th50', and 'petro_th90'. Since we want to discover 'how the galaxy's total size is related to the percentage of light within the radius', we will compare the three distribution graphs from the bootstrapping process in order to understand the relationship between the different sample statistics.

2.1.3 Visualization

The visualizations for research question 1 are three different `sampling distribution graphs` (which is a plot that shows the distribution of a statistics sample):

1. The first sampling distribution graph is related to the mean values of the 'petro_theta' statistic from the boot samples. We will set the x-axis to be the range of possible mean values from the boot samples, while we will set the y-axis to be the frequency (count) of a specific mean value.
2. The second sampling distribution graph is related to the mean values of the 'petro_th50' statistic from the boot samples. We will set the x-axis to be the range of possible mean values from the boot samples, while we will set the y-axis to be the frequency (count) of a specific mean value.
3. The third sampling distribution graph is related to the mean values of the 'petro_th90' statistic from the boot samples. We will set the x-axis to be the range of possible mean values from the boot samples, while we will set the y-axis to be the frequency (count) of a specific mean value.

The graphs are appropriate for the intended audience as sampling histograms were taught during lectures and are an easy way of visualizing the data. In addition to that, since we want to discover 'how the galaxy's total size is related to the percentage of light within the radius', we will compare these three distribution graphs in order to understand the relationship between the different sample statistics we are analyzing ('petro_theta', 'petro_th50', and 'petro_th90').

2.2 RQ2: Do the galaxies farthest to the Earth differ significantly from the closest ones in terms of their total luminosity?

H_0 = The mean of the total luminosity of the closest galaxies is not significantly different than that of the farthest ones.

H_A = The mean of the total luminosity of the closest galaxies is significantly different than that of the farthest ones.

2.2.1 Data

The data we will be using in research question 2 is from the ‘Galaxy Zoo Tabular Data Contents’. The variables involved in investigating the given question are [elpetro_absmag_r](#), which is the estimate of the total luminosity or, in other words, intrinsic brightness of the galaxies and [redshift](#) - a measure displaying how far away the galaxy is from us. Since our question aims to address two main attributes of the galaxies, total luminosity and distance from the Earth, the aforementioned variables are relevant.

2.2.2 Method

Our method for the second research question will be [test hypothesis](#). Our null hypothesis is ‘the difference of the means of the total luminosity values of the galaxies in the first and second samples being equal (or close to equal)’ while the alternative hypothesis is that there is a noticeable difference between the means. We will first look into the question of whether there is indeed a difference between the means of the total luminosities between the two samples, or not, and get our measured test statistic. Following this, we’ll use our null hypothesis to determine the probability options. Our test statistic will be the difference between the two means, which we will use to simulate the sampling distribution of that test statistic under the null hypothesis, e.g. 2000 times, with replacement. Then, we’ll graph the plot of the distribution based on our simulation results and calculate the 2-sided p-value according to the measured test statistic from earlier. Based on the rejection rule, we’ll either reject or fail to reject our null hypothesis depending on whether or not the p-value falls under the pre-specified alpha threshold (significance level).

2.2.3 Visualization

The visualizations for this research questions will be a [bar graph](#) and [histogram](#).

The former will be used to compare the actual means of the total luminosity values between the given samples, with its x-axis ‘which bar belongs to which sample’ and its y-axis showing the numerical mean value. Since there will be two overall samples, the plot will consist of two charts each mapping to a specific value of mean on y-axis. In this scenario, this plot type is the best choice, as the x-axis will contain discrete values rather than continuous.

We’ll also use histogram to visualize the distribution of the mean values we get from the sampling simulations under the null hypothesis. The choice of a histogram as our plot type for this part is appropriate since the x-axis will denote the possible mean values that will be produced as a result of the sampling simulations, which is a continuous numerical data, and y-value will represent their frequencies, i.e. the number of times the mean values in each interval appear upon the simulations. Histogram is also very comfortable in the sense that we will be

able to adjust the bin numbers of the graph (narrow down the intervals) depending on how detailed we want to convey the information to the reader.

2.3 RQ3: How well can a simple linear regression model predict the galaxy's apparent brightness from redshift?

2.3.1 Data

The data we used in research question 3 comes from 'Galaxy Zoo Tabular Data Contents'. They're [redshift](#), which is related to how far away that galaxy is from us, and, [seraic_nmygy_r](#), which is an estimation used to estimate the galaxy's apparent brightness (which depends on how far away it is) in units of magnitudes.

According to the description of redshift and [seraic_nmygy_r](#) above, 'galaxy's apparent brightness' in the research question 3 is revealed using the [seraic_nmygy_r](#) data and the 'redshift' in the research question 3 is revealed using the redshift data.

2.3.2 Method

The method we used in research question 3 is [simple linear regression](#). Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous or quantitative variables. To determine how well it relates, we'll estimate the regression line and then use the t-test to determine if the slope, β_1 , of the population regression line is 0. In relation to research question 3, since we want to know 'how well a simple linear regression model can predict the galaxy's apparent brightness from redshift', we set the 'redshift' as the predictor variable and 'the galaxy's apparent brightness' as the response variable, using simple linear regression method described above to estimate the regression line and apply t-test to evaluate the wellness of relationship.

2.3.3 Visualization

The visualizations for research question 3 are the [histogram](#) and the [fitted line plot](#).

For the [fitted line plot](#), it is generally used when obtaining the estimated regression function between a response y and a predictor x , which is as well exactly matches the case in research question 3 as we set 'redshift' as the predictor variable and 'the galaxy's apparent brightness' as the response variable. Therefore, in the fitted line plot, we'll set the x-axis as 'redshift' and the y-axis as 'the galaxy's apparent brightness', which is the same set as in the scatterplot. To justify the appropriation, although we already used the scatterplot graph to illustrate the dependent variables, the fitted line plot besides contains the shows a regression line superimposed on the data, providing much more accuracy when solving the research question 3, meaning it's appropriate.

For the [histogram](#), it is often used to illustrate the major features of the distribution of the data in a convenient form. Therefore, it is exactly the case in research question 3 as we set 'redshift' as x-axis in one histogram and 'the galaxy's apparent brightness' as another x-axis in one histogram. Since they are continuous variables, therefore, in the histogram, we'll compare the shape to illustrate their main feature by each of the histogram. To justify the appropriation, since we expect to get the main feature of those two data, it successfully satisfies the circumstances when using the histogram, which means it's appropriate.

3 Contribution

After team discussion, we regard the research question 3 the hardest research question. To ensure that the contributions of each group member is roughly an equal share to the overall workload, besides assigning a part of the whole procedure, we also assign each person a main research question to focus on.

1. Xuanqi Wei (RQ3): Conceptualization, Methods Determination, Coding;
2. Shujun Yang (RQ3): Data Cleaning, Introduction Section;
3. Riyad Valiyev (RQ2): Visualization, Conclusions Summarizing;
4. Nicolas Dias Martins (RQ1): Result Summarizing, Reviewing and Editing.

4 Timeline

1. By end of Feb: Submit the proposal for project. Each group member will choose their own questions to work on and share with other group members about their research questions as well as their dataset, methods, and visualizations they will use. The research questions will be accepted only when every group member understands all the concepts of the research questions, and agree it is a meaningful question.
2. By mid-Mar: Each member should try to finish the main part (coding, analysis) of their own research questions. A small meeting and frequent conversation will help with all the problems we faced during the process, the small meeting will help us to share our experiences and tell the difference between expectation and reality and how we deal with that so that we can make sure everyone is on the same page and all the adjustments we make are reasonable.
3. By Mar 30: All the members should finish their work and come up with a conclusion, each of us should send the related slides to one of our members so we can get a completed power point. The whole report should be done and check by all the group members even by TA. Two meetings will happened during this time, the first one is for the connection of reports and slides, the second one is the practice for presentation.
4. By Apr 5: All the work should be done and ready to submit. All the problems should be fixed by conversation in group and the help from TA.