Published April 11, 2023

Galaxy Zoo Data: Exploring and Analyzing Properties of Galaxies Using Hypothesis Testing and Bootstrap Sampling

Xufei (Annabella) An, Betty Li, Haochen (Thomas) Li, and Edric Shi

Abstract

The Galaxy Zoo Data provided by astrophysicist Mike Walmsley and NASA-Sloan Atlas contains information on a fraction of roughly one hundred billion galaxies around the observed universe (Zooniverse, 2007). The goal of the Galaxy Zoo Project is to understand the properties and ongoing changes in these galaxies.

In this report, we will discover various properties of the galaxies in the Galaxy10 dataset, including the redshift, sizes, and apparent brightness of the galaxies (NASA-Sloan Atlas). Through hypothesis testing and bootstrap confidence interval testing, we will investigate the mean and median of these properties, and conclude what they could mean to the research. After a series of data analyses were performed, we have found that the true mean of the galaxies' redshift is not equal to 0.08, size medians are different for small and large galaxies, and the medians of galaxies' apparent brightness using sersic_nmygy_r and mag_r data are different.

Our research and data analysis not only inherited the aspiration of the original project but also paved a path for potential future work as we have used standardized methods in data analysis to perform an investigation on a small portion of this large dataset. We hope to gain a deeper understanding of the universe through our fruitful conclusions on the numerous galaxies around us.

Introduction

Outer space has been a mystery to human beings since the beginning of time. Its complexity, its chaos, and its vitality attract thousands of astrophysicists to an attempt to explore the wonder. The Galaxy Zoo Project, which aims to perform deeper research on the eyes of the universe —— the galaxies, believes we could use what we understand about the galaxies to conclude on the past, present, and future of the universe as a whole. Statistics and data analysis along with galaxy images are used as tools for us to explore the universe and the galaxies. Until today, under the effort of a number of astrophysicists and researchers and about 10,000 volunteers in the team, 32% of the Galaxy Zoo statistics are completed for further analysis.

Our research group of four members was impressed by this huge accomplishment and decided to contribute to the project through a series of data analyses on the basic properties of the galaxies in the dataset. We have developed three research questions and had five variables involved in our report with the hope to investigate the properties of the galaxies. Furthermore, statistical methods of hypothesis testing and bootstrap confidence interval testing were used for effective data analysis because they were perfect methods to deal with such a huge dataset of galaxies.

Data

```
library(arrow)
library(tidyverse)
```

```
df <- read_parquet("nsa_v1_0_1_key_cols.parquet")
glimpse(df)
```

```
## Rows: 641,409
## Columns: 10
## $ ra              <dbl> 146.7142, 146.6286, 146.6317, 1
## $ dec             <dbl> -1.04128002, -0.76516210, -0.98
## $ iauname         <chr> "J094651.40-010228.5", "J09463(
## $ petro_theta     <dbl> 7.247893, 5.617822, 4.769891, 6
## $ petro_th50      <dbl> 3.464192, 2.326989, 2.278736, 2
## $ petro_th90      <dbl> 10.453795, 6.721991, 5.177910,
## $ elpetro_absmag_r <dbl> -19.30366, -19.97650, -18.4318:
## $ sersic_nmgy_r   <dbl> 1789.25720, 229.84039, 82.22815
## $ redshift        <dbl> 0.021222278, 0.064656317, 0.052
## $ mag_r           <dbl> 14.36832, 16.59643, 17.71245, 1
```

Research Question 1

Question

is the true mean of galaxies' redshift 0.08?

Data

The variable used for this research question is "redshift" in the data set.

Method

By only observation, with no calculation of the data "redshift", the mean appears to be approximately 0.08. We set the Null Hypothesis as the mean of redshift equal to 0.08, and the Alternative Hypothesis as the mean of redshift not equal to 0.08.

To evaluate the data we used the "summarise()" function to find the mean, median, max, and standard deviation. Then use the functions "ggplot()", "aes()", and "geom_boxplot()" to visualize the data.

Analysis

After returning the function we found out that the mean = 0.08518334, median = 0.08487365, max = 0.1499996, and the standard deviation = 0.03650291. The boxplot had the same result where the median is around 0.085.

Results

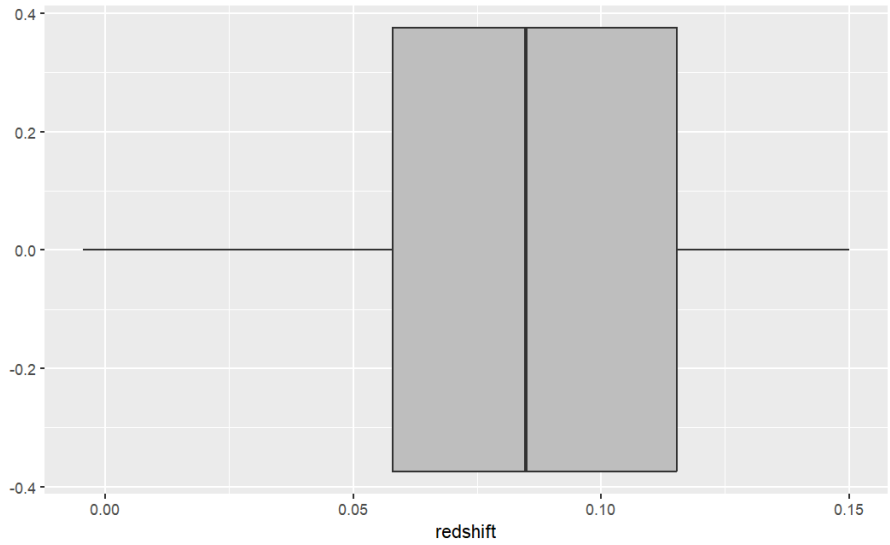
The results of the "summarise()" function showed that the mean = 0.08518334, which is strong evidence against the Null Hypothesis, so the Alternative Hypothesis is more accurate. The answer to the research question is no, the true mean of galaxies' redshift is not 0.08.

Data Visualizations

A tibble: 1 × 4

mean <dbl>	median <dbl>	max <dbl>	sd <dbl>
0.08518334	0.08487365	0.1499996	0.03650291

1 row



Research Question 2

Question

What is the 95% confidence interval of the estimate of the galaxy's size where 50% of the light is within the radius and 50% of the light is outside of the radius of small and large galaxies are the same?

Data

The variable that has been chosen for this question is ‘petro_th50'. 'petro_th50' represents an alternative estimate of the galaxy's size where 50% of the light is within the radius and 50% of the light is outside of the radius, demonstrated by its radius.

In the progress of cleaning data, we created a new variable called ***size*** to split the sizes in petro_th50 into “small” and “large” to create two samples for our two-sample hypothesis testing. The sizes that are smaller or equal to the 50th percentile of the set are considered “small,” and the rest of the sizes are considered ‘’large.” Then we used the filter function to eliminate the empty values in the data set.

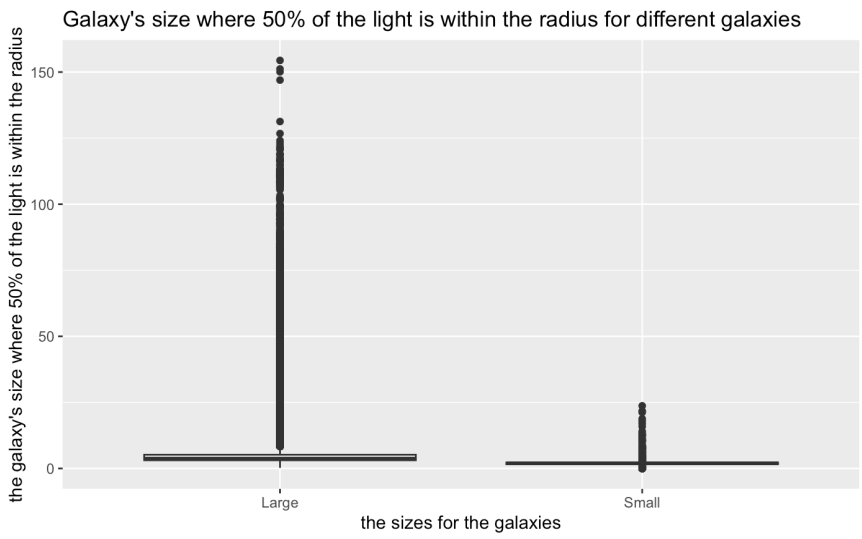
Method

We choose to use two-sample hypothesis testing for several reasons: firstly, our purpose of the test is to determine whether the difference between large and small galaxies is statistically significant; second, our data size is large enough thus, we do not need to use the bootstrap method. Therefore, the two-sample hypothesis testing best suits our purpose.

In the two-sample hypothesis testing, the null hypothesis is $$H_0: M_{small} = M_{large}$$; the alternative hypothesis is $$H_0: M_{small} != M_{large}$$, where M_{small} is the median for an alternative estimate of the galaxy's size where 50% of the light is within the radius, and 50% of the light is outside of the radius of small galaxies, and M_{large} is median for an alternative estimate of the larger galaxies.

Data Visualization

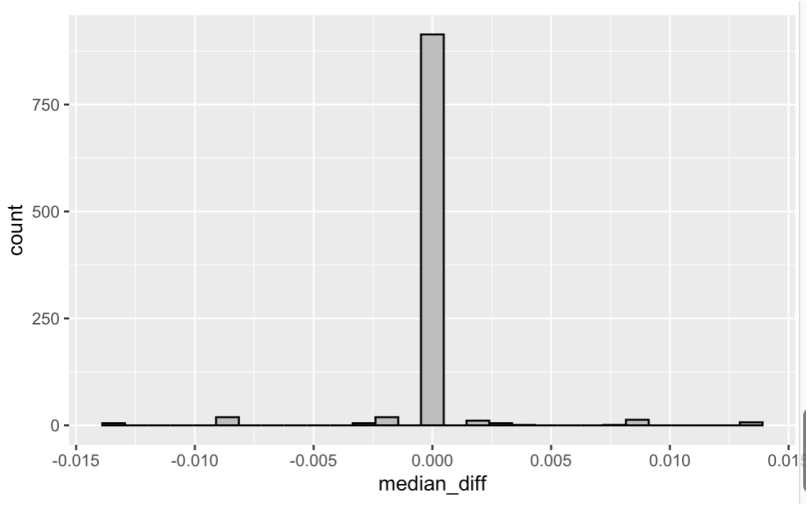
We used a boxplot to take a preliminary visualization of the data set because the boxplot best demonstrates the median, which is also known as the 50th percentile, of the galaxies’ sizes. The boxplot below shows the data distribution of small and large galaxies’ sizes. Its x-axis is the size categories of the galaxies, and the y-axis is its precise radius.



Analysis & Results

Getting into the data analysis part, we first calculated the test statistics of the median difference of the small and large galaxies by using the 'summarise' function, which gives us a number of -1.85 (rounded to 2 decimal places).

In the next step, we did the two-sample hypothesis testing by first shuffling the group, then grouping by ***size*** to calculate the median difference of petro_th50 of small size group and large size group. We repeated the process 1,000 times and store all the simulated sample median differences in an empty list with a size of 1,000. The set.seed( ) function helps us to create reproducible results. Then we use the list containing all the median differences to make the histogram shown below. From the histogram, we could see that the distribution is unimodal centred at 0.



Then we calculated the p-value based on the simulated data and test statistics we got from the previous steps by dividing the number that is more extreme by the number of repetitions. We got a p-value of 0 from the calculations.

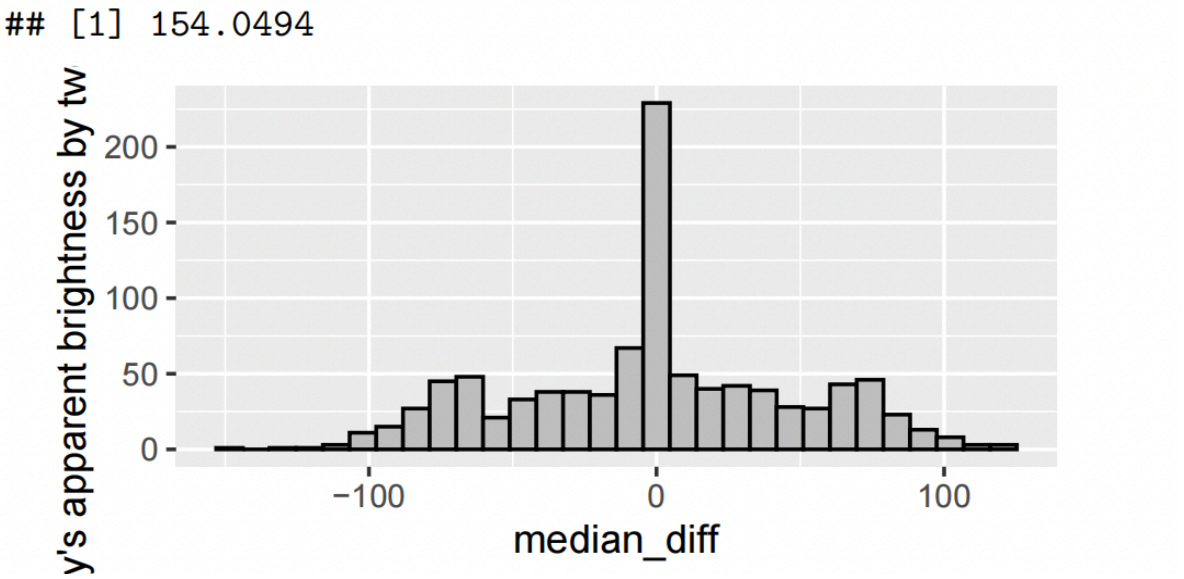
In the last step, we calculated the 95% confidence interval of the difference median using the quantile function to find the quantile from 0.025 to 0.975. We finally got the confidence interval is -0.0029 to 0.0029.

According to the p-value of 0, we could conclude that we have very strong evidence against the null hypothesis. The result of the confidence interval demonstrates that we have 95% of confidence to say that the true median difference between small and large galaxies' sizes is in the interval from -0.0029 to 0.0029.

Research Question 3

The third research question explores whether the medians of the galaxies' apparent brightness using sersic_nmygy_r data and the galaxies' apparent brightness using mag_r data are the same.

Data Visualization



```
## # A tibble: 1 x 1
##   pvalue
##    <dbl>
## 1      0
```

Method & Analysis

In this research question, we used a two-sample hypothesis test for the median. Both data can be found in the data Galaxy10, and while selecting them, we subsetted a data frame and retained the rows that satisfy our conditions by using filter( ) function. We first defined its null hypothesis and alternative hypothesis, the former is that the two data groups have the same median, while the latter is that they have different medians. For this test, we randomly divide the combined dataset into two groups by using set.seed( ) function, which is a method to ensure that the results of random processes are reproducible because the same set of random numbers will be generated each time the function is run with the same seed value. For the results of the whole data set, we set the same sample sizes as the original groups, which is 641409. Then, we calculated the median for each of the randomly divided groups, followed by the calculation of the difference between the two medians by setting a test_stat. Then, after shuffling the groups and computing the median difference, we construct an empty list and add the simulated sample median difference to it. To obtain a distribution of the median differences under the null hypothesis, we repeated these steps 1000 times. For further comparison, we plotted a histogram to compare the observed median difference from the original two groups with the distribution of median differences obtained from the shuffling process. Lastly, to determine the p-value, which is the proportion of times that the median difference from the shuffling process was as extreme or more extreme than the observed median difference, we divided the simulated median differences we processed that were larger than our test statistics by the number of observations, which is 1000 times.

Results

In the end, we got a test_stat of 154.0494, which is the difference between the median of sersic_nmygy_r data and the one of mag_r data. Based on the results of the previous analysis, we also got a significantly small p-value of 0. Since the p-value is smaller than 0.001, it is obvious that there is very strong evidence against the null hypothesis. Therefore, the result of this question is that the medians of the galaxies' apparent brightness using sersic_nmygy_r data and the galaxies' apparent brightness using mag_r data are not the same. Such a two-sample hypothesis test can provide valuable information based on the sersic_nmygy_r data and mag_r data provided in Galaxy10, and can help us draw conclusions on the differences between two methods of estimating galaxies' apparent brightness in units of magnitude.

Conclusion

As stated before, this report, which was inspired by the Galaxy Zoo project and curiosity toward space and the universe, contains exploration and data analysis on the properties of the galaxies in the Galaxy10 dataset (Zooniverse, 2007). Our research team has used hypothesis testing and bootstrap confidence interval testing for determining the means and medians of the redshift, sizes, and apparent brightness of the galaxies (NASA-Sloan Atlas). We have hence concluded that the true mean of the galaxies' redshift is not equal to 0.08, size medians are different for small and large galaxies, and the medians of galaxies' apparent brightness using sersic_nmygy_r and mag_r data are different. These results provide tons of implications to researchers and scholars in astronomy and astrophysicist as they could apply the results in further investigations of the galaxies and the universe.

Our research and data analysis not only inherited the aspiration of the original project but also paved a path for potential future work as we have used standardized methods in data analysis to perform an investigation on a small portion of this large dataset. We believe further research into the galaxies provided by the Galaxy Zoo data would enhance our understanding of the broader universe that every one of us lives in.

Citations

1. Howell, E. & Dobrijevic, D. (2022, January 14). *Redshift and blueshift: What do they mean?*. Space. https://www.space.com/25732-redshift-blueshift.html
2. NASA-Sloan Atlas. (n.d.). *Datamodel: nsa*. https://data.sdss.org/datamodel/files/ATLAS_DATA/ATLAS_MAJOR_VERSION/nsa.html
3. Wikipedia contributors. (2023, March 29). *Galaxy Zoo*. In Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Galaxy_Zoo&oldid=1147249389
4. Zooniverse. (2007, July 11). *Galaxy Zoo*. https://www.zooniverse.org/projects/zookeeper/galaxy-zoo

Appendix

Throughout the process, we encountered some questions that could not be solved or results that could not be explained. The first main one was that research question 2 was changed from finding the difference in median of the apparent brightness of small and large galaxies to determining whether they are equal or not, because the former could not be implemented and answered using current knowledge from STA130. The second concern was the resulting p-value of 0 in research question 3, which was not possible. By our conjecture, the actual p-value was so small that R Studio rounded it off to 0. Although the two concerns could be justified with logical reasons, it is important to address them here to resolve any potential confusion in the study.