Capstone Project

---- Use Two-sample Hypothesis Test, Multivariate Linear Regression Model, and Bootstrap Method to Analyze Data in the Galaxy Zoo dataset

Peize Zhang Raghav Sinha
Yuhang Yang Jiaxun zhang

Abstract:
This project contains three research questions about a characteristic of galaxies, redshift. Various statistical methods have been used. Data cleaning and preparation, correlation analysis using tools like correlation matrices, boxplots, histograms and scatterplots. And finally, two-sample hypothesis test, regression analysis using both robust and non-robust multivariate and bootstrap methods to analyze these questions. Finally, we figured out there is no evidence against the redshift of class 1 and class 2 Galaxies are same, there is a correlation between luminosity and its redshift, and we have 95% to say that the true mean of the redshift of class four within (0.05281380, 0.08019046). These results can be utilized for future research on galaxy information, and we compared them with other galaxies to identify similarities and differences in their redshift characteristics.

■ Research question 1

Is the mean distance(redshift) from Merging Galaxies(class 1) and Round Smooth Galaxies(class 2) to the Earth the same?

Introduction:

Scientists have classified galaxies into different classes. I'm curious about if redshift value is one of the bases of classification. This sparks me to check if the mean redshift of two galaxies are the same using what I learned in STA130. Redshift is a quantitative variable measuring the distance from a galaxy to Earth. I used two sample hypothesis test to solve the problem. This is a statistical method used to test if the two unknown population means are the same or not. In my research, both graphs and tests are used in order to present the result more clearly and accurately.

Data:

I used the data set called "Galaxy10_DECals.h5", which contains galaxies of all classes and their corresponding redshift values. I first used as_tibble() function to assemble all the data into a table. This enables me to extract and analyze the data more easily.

In the data cleaning process, I used filter() function to extract those class 1 and class 2 galaxies and their corresponding redshift values. However, some of the redshift values are NA values, which means we haven't yet measured the distance of those galaxies from Earth. Facing this situation, I use !is.na() function to remove those NA values.

Below are the detailed codes for data cleaning:

```r
redshift = "Galaxy10_DECals.h5" %>% h5read("redshift") %>% as_tibble()
galaxy_class = "Galaxy10_DECals.h5" %>% h5read("ans") %>% as.integer() %>%
as.factor() %>% as_tibble()
```

```r
ana_data= galaxy_data %>% mutate(across(galaxy_class, as.character)) %>%
filter(galaxy_class == "1" | galaxy_class == "2", !is.na(redshift))
ana_data
```

A tibble: 4,483 × 2

galaxy_class <chr>	redshift <dbl>
1	0.134631100
1	0.121170126
1	0.098265570
1	0.081034176
1	0.098783955
1	0.090725990
1	0.132068110
1	0.109501235
1	0.148849550
1	0.139103350

1-10 of 4,483 rows Previous 1 2 3 4 5 6 ... 100 Next

Methods/Analysis:

I used two sample hypothesis test to compare the mean redshifts between class 1 and class 2 galaxies. I began with the null and alternative hypothesis:

* null hypothesis: the mean redshift between class 1 and class 2 galaxies are the same.
* alternative hypothesis: the mean redshift between class 1 and class 2 galaxies are different.

Here we assume the null hypothesis is true.

With data of two galaxy classes in hand, I first used group_by() function to split the galaxies into two groups based on their classes. In each group, one sample redshift is randomly selected. I calculated the difference between the two redshift values. The output of the process is the test statistics.

Next, I repeated the above process for 1000 times using for loop. After each repetition, I shuffled all the data and split them up again. This makes sure each sample is selected randomly. Then, I made difference of each sample pairs and stored the output for further analysis.

Finally, by comparing each simulated sample difference with the test statistics, I used sum() function to generate the p-value which is used to judge if our assumption is true.

```r
repetition = 1000
sim_dif = rep(NA, repetition)
for(i in 1: repetition){
  difference = ana_data %>% mutate(galaxy_class = sample(galaxy_class)) %>%
group_by(galaxy_class) %>% summarise(means = mean(redshift)) %>%
summarise(diff(means)) %>% as.numeric()
  sim_dif[i] = difference
}
```

```r
p_value = sum(sim_dif > test_sta)/repetition + sum(sim_dif < -1 * test_sta )/
repetition
p_value
```

```
[1] 0.512
```

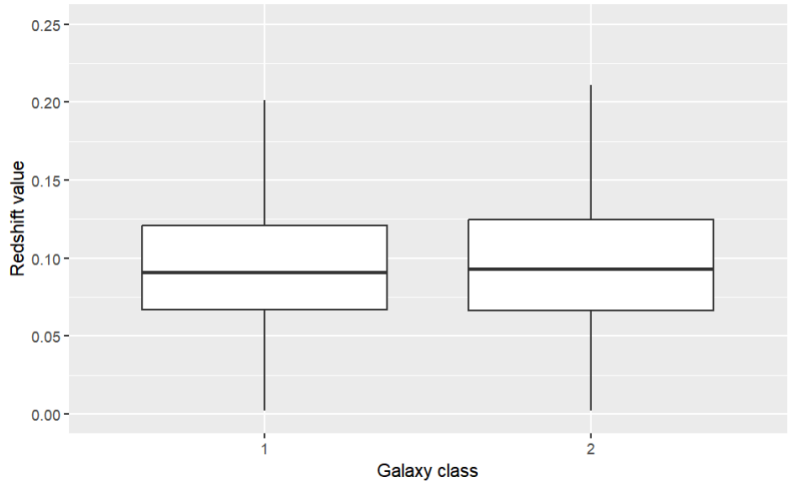
Our p-value is 0.512. While the significant level is 0.05. The significant level is the strength of evidence that must be present in our sample before rejecting the null hypothesis. Here 0.512 is much larger than 0.05. It means that we don't have enough evidence to reject the null hypothesis.

Results:

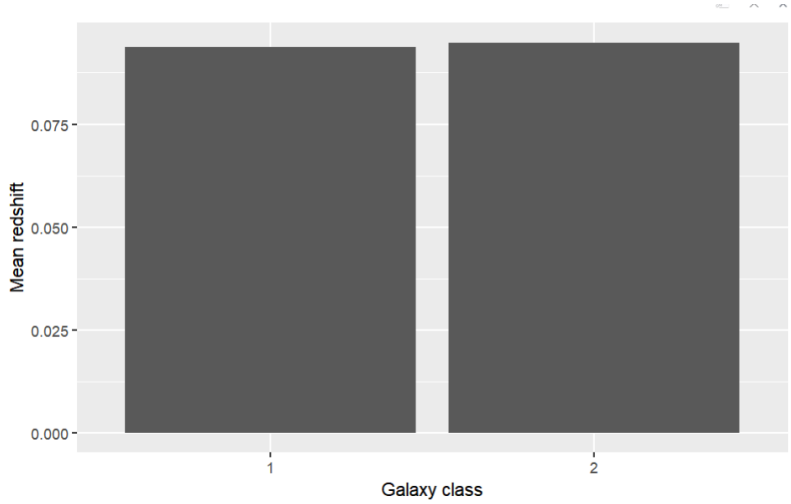
Our large p-value tells that we don't have enough evidence to reject the null hypothesis. It means we don't have enough evidence to say the mean redshifts of two galaxy classes are different.

Data visualizations:

Boxplot showing the general distribution of redshifts of two galaxy classes:



Barplot comparing the mean redshift of two galaxy classes:



Discussion:
The two sample hypothesis test successfully helps to solve the problem. Result shows that we don't have enough evidence to reject the null hypothesis. However, failing to reject the null hypothesis doesn't mean the null hypothesis must be true. It just represents null hypothesis is not false.

In addition, what I did was a test. It means all the output data is simulated data used to reflect what the real data might be. So we're actually making an approximation and there are definitely errors occur. In reality, we don't really know what the mean redshift difference between the two galaxy classes is.

Conclusion:
In the research, I used two sample hypothesis test to check if the mean redshift of class 1 and class 2 galaxies are the same. I did all these in order to see if redshift value is one of the bases of galaxy classification. So what we can get from the test is we can't say the mean redshift of class 1 and class 2 galaxies are different based on our research. As a result, our conclusion is we don't have enough evidence that redshift is a basis for galaxy classification.

Citation:
Google search: two sample hypothesis test:
https://en.wikipedia.org/wiki/Two-sample_hypothesis_testing

■ Research Question 2

What is the correlation, if any, between the distance to a galaxy (measured by redshift) and its luminosity in absolute magnitudes (measured by elpetro_absmag_r)?

Title:
Investigating the Relationship Between Redshift and Galaxy Luminosity using Multivariate Linear Regression

Introduction:
Galaxies, and most other objects in the Galaxy Zoo dataset were observed using

telescopes. Contrary to popular belief, most of this data is numerical (not optical). The variables relevant to us are ‘redshift’, ‘elpetro_absmag_r’, and ‘mag_r’.

i. **redshift** is a measure of how much the light from an object (such as a galaxy) has been stretched or "shifted" towards the red end of the spectrum due to the object's motion away from us. In simpler terms, it tells us how far away a galaxy is from Earth.
ii. **elpetro_absmag_r** is a measure of the absolute magnitude of a galaxy in the r-band of the electromagnetic spectrum. Magnitude is a measure of the brightness of an object, and "absolute" magnitude is a way of comparing the brightness of objects at a fixed distance (in this case, 10 parsecs).
iii. **mag_r** is a measure of the apparent magnitude of a galaxy in the r-band of the electromagnetic spectrum. Apparent magnitude is a measure of how bright an object appears to be from Earth, without taking into account its distance from us.

We are trying to find correlation between either elpetro_absmag_r and redshift or mag_r and redshift. This will be done using linear regression.

Data:
I am using the Galaxy Zoo NSA dataset – “nsa_v1_0_1_key_cols.parquet” – after performing a series of cleaning operations.
The file is in the parquet format. To use this as a dataframe, I used the ‘read_parquet()’ function from the ‘arrow’ package.

```r
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, 146.9341, 146.9635, 146.9635, 146.8598, 146.5928, 146.7284, 146.6072, 147.1868,…
$ dec             <dbl> -1.04128002, -0.76516210, -0.98834670, -0.67040536, -0.54477583, -0.75934042, -0.80890650, -0.76025740, -0.55…
$ iauname         <chr> "J094651.40-010228.5", "J094630.85-004554.5", "J094631.59-005917.7", "J094744.18-004013.4", "J094751.74-00324…
$ petro_theta     <dbl> 7.247893, 5.617822, 4.769891, 6.243227, 8.891541, 3.549432, 5.860092, 5.953613, 43.310619, 10.695449, 13.5242…
$ petro_th50      <dbl> 3.4641922, 2.3269887, 2.2787361, 2.6551907, 4.3837042, 1.6775934, 2.6134822, 2.7990863, 16.4872322, 4.3376756…
$ petro_th90      <dbl> 10.4537954, 6.7219906, 5.1779103, 9.1776047, 10.2532740, 4.8975000, 7.9587302, 7.6296973, 37.5585480, 14.2223…
$ elpetro_absmag_r <dbl> -19.30366, -19.97650, -18.43181, -21.55916, -19.10099, -20.08994, -21.06252, -19.97614, -19.90020, -21.19160,…
$ sersic_nmgy_r   <dbl> 1789.25720, 229.84039, 82.22815, 277.76120, 132.77216, 120.70260, 124.92822, 239.27782, 121.79979, 455.03537,…
$ redshift        <dbl> 0.021222278, 0.064656317, 0.052654251, 0.121270485, 0.055980586, 0.097086377, 0.126589879, 0.064959235, 0.089…
$ mag_r           <dbl> 14.36832, 16.59643, 17.71245, 16.39082, 17.19223, 17.29571, 17.25835, 16.55274, 17.28588, 15.85489, 16.18922,…
```

We see quite a range of variables in this dataset. For us, the relevant columns are elpetro_absmag_r, redshift and possibly mag_r

```r
df_numeric <- df %>% dplyr::select(-c("ra", "dec", "iauname"))
df_numeric <- df_numeric %>% mutate(log_redshift = log10(redshift))

df_numeric <- na.omit(df_numeric)
df_numeric <- df_numeric[!is.infinite(rowSums(df_numeric)),]
```

This was read in and then the NA, inf and NaN values were removed from the dataset. Additionally, after observing the scatterplot between redshift and elpetro_absmag_r, I decided to create a new column called log_redshift.
This was done to linearize the relationship so linear regression would be more successful.

Note that since we have imported both ‘tidyverse’ and ‘MASS’ I had to explicitly specify that the select function was from the dplyr package from tidyverse.

Methods/Analysis:
Once we're done cleaning and pruning the data, we can get to the actual analysis.
First, I formulated my null and alternative hypotheses:

> Null hypothesis: There is no significant relationship between mag_r + elpetro_absmag_r and log_redshift.
>
> Alternative hypothesis: There is a significant relationship between mag_r + elpetro_absmag_r and log_redshift.

As we are trying to find correlation between our variables, it's always a good idea to generate a correlation matrix of the dataset. I did this using the 'corrplot' package.

```r
corr <- cor(df_numeric)

# Create a correlation matrix heatmap
corrplot(corr, type = "upper", method = "color",
         tl.col = "black", tl.srt = 45,
         addCoef.col = "black",
         number.cex = 0.8, order = "hclust", title="Dataset Correlation Matrix Heatmap")
```

This generates a correlation matrix between each variable in the data and then plots it as a heatmap.
From this, we can confirm our selected variables and observe other interesting relationships!

I observed that there is a strong negative correlation between log_redshift and mag_r as well as log_redshift and elpetro_absmag_r.
This meant I was on the right path. The next step is to visualize this correlation using a scatterplot.

```r
ggplot(df_numeric, aes(x=log_redshift, y=elpetro_absmag_r)) +
  geom_point(color="blue", alpha=0.5, size=2) +
  geom_smooth(method="lm", se=FALSE, color="red", size=1) +

  # Set the title and axis labels
  labs(title="Scatterplot of log_redshift vs elpetro_absmag_r",
       x="log_redshift",
       y="elpetro_absmag_r")
```

I confirmed that the sigma value for the robust regression was lower than that of the non-robust regression:

```r
mod_redshift_mag <- lm(log_redshift ~ mag_r+elpetro_absmag_r, data = df_numeric)
cat("sigma value for non-robust linear regression:", summary(mod_redshift_mag)$sigma, "\n")

mod_redshift_mag_r <- rlm(log_redshift ~ mag_r+elpetro_absmag_r, data = df_numeric)
cat("sigma value for robust linear regression:", summary(mod_redshift_mag_r)$sigma, "\n")
```

```
sigma value for non-robust linear regression: 0.08081461
sigma value for robust linear regression: 0.02097261
```

This makes sense as I observed quite a few outliers from the boxplot I generated. The robust linear regression is the better option.

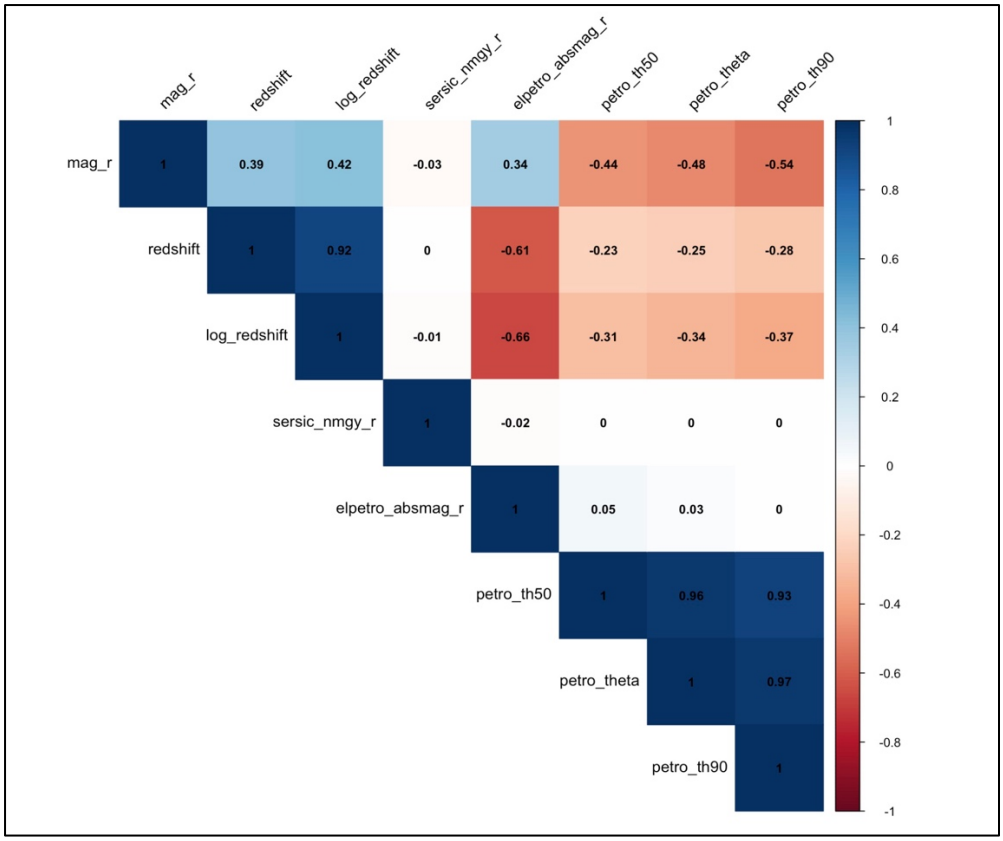
The final step is the regression analysis itself. After looking at the scatterplot, I decided that a robust linear regression would be better because of the large number of outliers.
For this, I used the 'rlm()' function from the 'MASS' package.
The following summary table was produced by the robust linear regression model:

```
ROBUST LINEAR REGRESSION ANALYSIS

Call: rlm(formula = log_redshift ~ mag_r + elpetro_absmag_r, data = df_numeric)
Residuals:
     Min       1Q   Median       3Q      Max
-2.95530 -0.01394 -0.00120  0.01445  3.11860

Coefficients:
                Value       Std. Error  t value
(Intercept)          -7.8859      0.0007 -10668.9859
mag_r                 0.1764      0.0000   6702.2214
elpetro_absmag_r     -0.1891      0.0000  -8228.4094

Residual standard error: 0.02097 on 640770 degrees of freedom
```

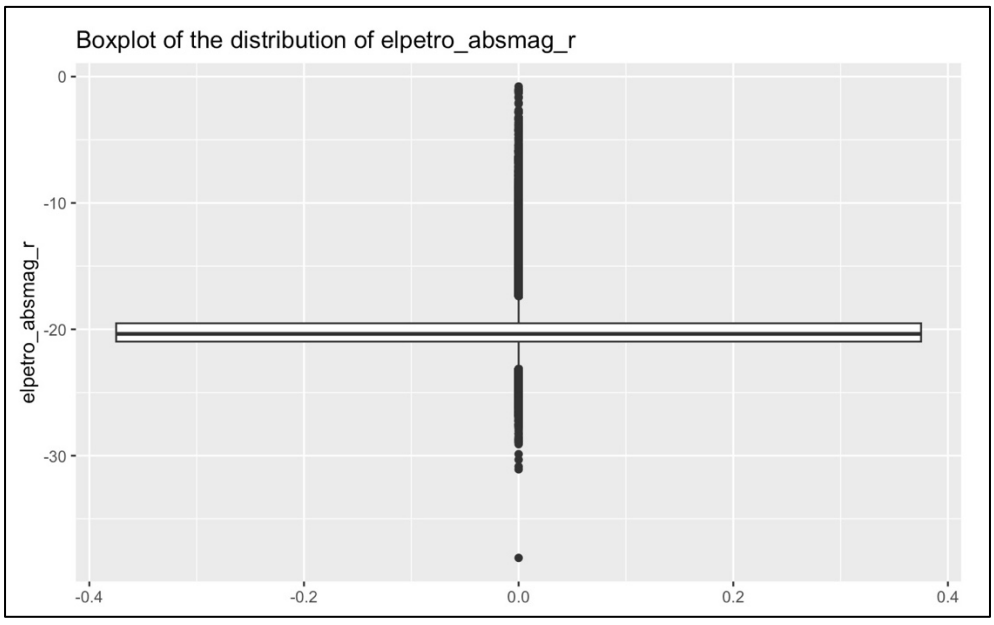
Results:
From the regression model, we can see that the standard errors for the coefficients are very small, indicating a high level of precision in the estimates. Additionally, the t-values for both coefficients are very large, indicating that the coefficients are statistically significant. (i.e. we can reject the null hypothesis).
We also see a very small standard error, which reinforces our results' statistical significance.

These results suggest that the brightness and absolute magnitude of a galaxy are important factors that influence its redshift. This is consistent with previous research in the field, which has shown that there is a correlation between the luminosity of a galaxy and its redshift.

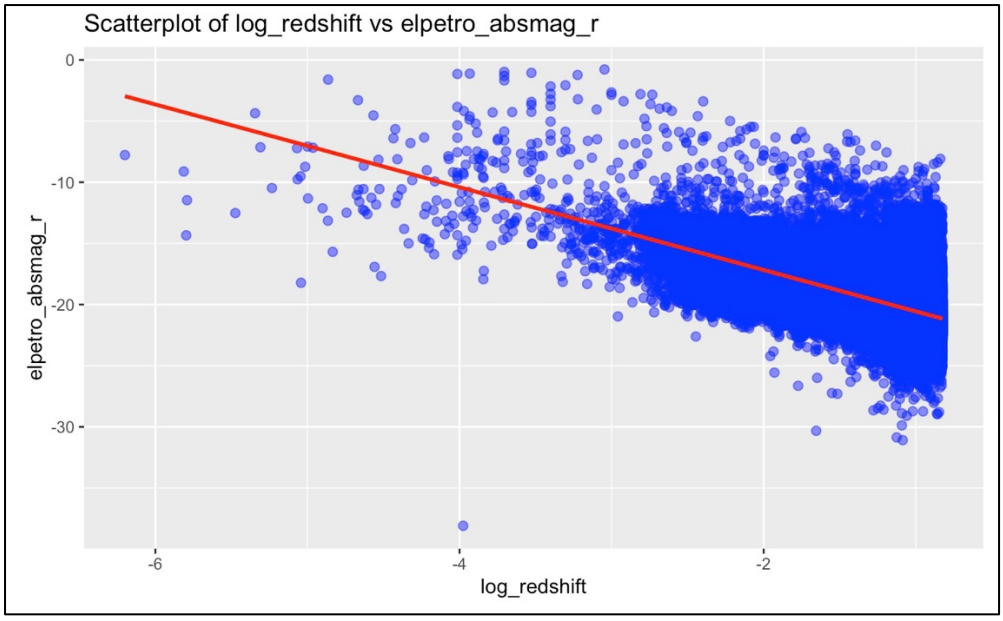
Data Visualizations:
Correlation Heatmap showing the correlation between each variable in the dataset:



Boxplot showing outliers:



Scatterplot of log_redshift vs elpetro_absmag_r:



Discussion:

The focus of this work was to determine the nature of correlation between the two variables redshift and luminosity. But what do they really imply? Putting this work in context with the popular galaxy luminosity function gives us some interesting ideas.

The galaxy luminosity function is a statistical distribution that describes the number of galaxies in the universe as a function of their intrinsic brightness (luminosity).
The luminosity function is an important tool in astronomy, as it provides a way to understand the distribution and evolution of galaxies in the universe. It can be used to study the formation and evolution of galaxies, the large-scale structure of the universe, and the properties of dark matter.
Observationally, the galaxy luminosity function is derived by counting the number of galaxies in a given volume of space and measuring their apparent brightnesses (fluxes). By correcting for the effects of distance, redshift, and survey completeness, astronomers can then construct a luminosity function that describes the true distribution of galaxy luminosities in the universe.

Thus, our correlation study was the first of many steps to deriving this quantity. Knowing how quantities relate is an integral part of expanding our understanding of the objects from numbers to the real world.

Conclusion:

In conclusion, we explored what redshift and magnitudes mean, got an idea of whether they are correlated or not and then performed a regression analysis to determine the nature and strength of their correlation as well as whether these two variables are actually influencing each other. In the end, we interpreted these results in the context of real world problems that astronomers are trying to solve. I would further like to consider the types of galaxies by combining the two given datasets so we can get a better idea of how the class of a galaxy influences this correlation.

Citations:

Salpeter, E E, and G L Hoffman. “The galaxy luminosity function and the redshift-distance controversy (A Review).” *Proceedings of the National Academy of Sciences of the United States of America* vol. 83,10 (1986): 3056-63. doi:10.1073/pnas.83.10.3056

■ Research question 3

What are the plausible values for the average redshift of Cigar-Shaped Smooth Galaxies?

Introduction:

The study of galaxies is crucial for understanding the formation and evolution of our universe. One of the key factors in this study is the redshift of galaxies, which provides valuable information about how far away it is from Earth and how quickly it is departing; a galaxy's redshift value tells us how near it is to Earth and how slowly it is departing. We can learn more about the distribution of Cigar-Shaped Smooth Galaxies in the cosmos and gain insight into their general properties by determining the true mean redshift of these galaxies. So, for this question, we estimate the average redshift of Cigar-Shaped Smooth Galaxies (Class 4) using a statistical method called bootstrap, which can resample group data to estimate the sampling distribution of that. By using the distribution of that, get a 95% confidence interval of the statistics.

Data:

For the third question, we address Galaxy10 data, using “Galaxy10_DECals.h5”. Selecting two variables called “ans” and “redshift. The “ans” is a categorical variable which contains 334 images of Class 4 (Cigar-Shaped Smooth Galaxies). This variable provides information about the Galaxy’s class. And the “redshift” is a numerical variable which measures the distance from the Earth. And we remove the NA data from this variable. So, these two variables allow us to measure all sample averages, then conclude the plausible values for the average redshift of Cigar-Shaped Smooth Galaxies.

Method:

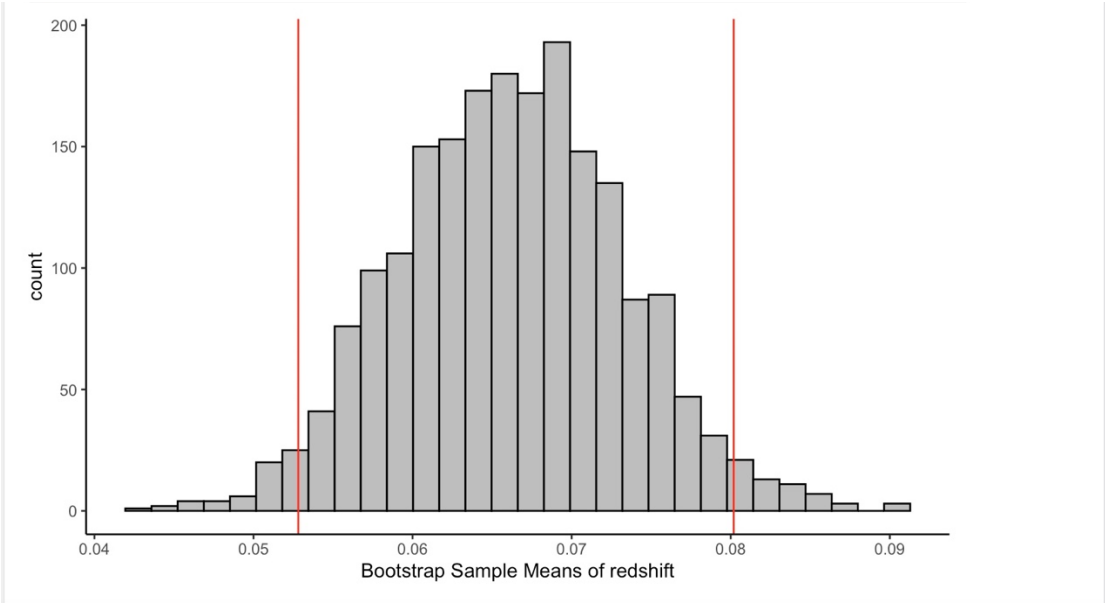
First, we summarize the redshift of Class 4: Cigar_Shaped Smooth Galxies, and make a boxplot to see the distribution of that. Then, we use the Bootstrap confidence intervals to evaluate question 3. Because all redshifts of Cigar-Shaped Smooth Galaxies, which are the population of this question, are unavailable, we can estimate the sampling distribution of the sample average, starting with a sample of 20 with observed test statistics average. Then resampling many times to draw a bootstrap sample of size 20, with replacement from the original sample, then for each bootstrap sample, calculate the statistic.

Therefore, we can get the distribution of bootstrap statistics. A 95% confidence interval for the parameter is the middle 95% value of the bootstrap statistics. This interval should be the plausible value on the true population average of the redshift of Cigar-Shaped Smooth Galaxies.

Result:
We have 95% confidence to show that the true mean of the redshift of Cigar-Shaped Smooth Galaxies is within the interval (0.05281380, 0.08019046).

Data visualizations:
Boxplot showing the general distribution of redshifts of Class 4: Cigar_Shaped Smooth Galaxies:

The distribution of bootstrap sample means of redshifts of Class 4: Cigar_Shaped Smooth Galaxies and 95% confident interval:



Discussion:
Our analysis of the Galaxy10 dataset has provided valuable insights into the redshift distribution of Cigar-Shaped Smooth Galaxies. The calculated 95% confidence

interval for the true mean redshift of these galaxies, (0.05281380, 0.08019046), offers a range of plausible values for this population parameter.

The results are statistically significant, as the bootstrap method used in the analysis accounts takes into account the sample data's variability and uncertainty and yields statistically significant results. We may be 95% confident. The confidence interval we derived indicates that We have 95% confidence that the true mean redshift of cigar-shaped smooth galaxies falls within the given range. This interval also allows us to make inferences about the underlying data without relying on assumptions about the distribution of redshift values in the population.

And we discover that when working with complex distributions, we can utilize Bootstrap to estimate population parameters. We created a confidence interval that takes into account the variability in the data without making any assumptions about the distribution's underlying properties by resampling several bootstrap samples and computing the corresponding sample means.

Conclusion:
In conclusion, our analysis of the Galaxy10 dataset has provided a 95% confidence interval for the true mean redshift of Cigar-Shaped Smooth Galaxies. The interval estimated in this study, (0.05281380, 0.08019046), gives us a range of plausible values for the average redshift of these galaxies. By using the bootstrap method, we were able to get the distribution of bootstrap statistics.

Compared to galaxies with greater redshift values, those with low redshift values are often closer to Earth and have not experienced as much expansion. Due to their low redshift range, cigar-shaped smooth galaxies are typically closer to us than some other galaxy shapes that might have higher redshift values.

This research question helps us better understand the characteristics of Cigar-Shaped Smooth Galaxies and their position in the universe relative to Earth, which can be utilized to inform future research on galaxy formation. And the results can be compared with other galaxy classes to identify similarities and differences in their redshift characteristics.

Citations:
https://www.space.com/25732-redshift-blueshift.html

Contribution:

	Data cleaning	Visualization and Summary table	Analysis and Writing	Integrate group work
Peize Zhang	Question 1	Question 1	Question 1	

Raghav Sinha	Question 2	Question 2	Question 2	
Yuhang Yang	Question 3	Question 3	Question 3	√
Jiaxun Zhang	Question 3	Question 3	Question 3	√

In our group, the individual contributions are roughly equal. Peize Zhang is responsible for question 1 work, Raghav Sinha works on question 2, and Jiaxun Zhang and Yuhang Yang are responsible for question 3. Each question's work includes pre-processing and cleaning data, visualizing and summarizing tables, and writing analysis tasks. Finally, Yuhang Yang and Jiaxun Zhang will integrate all group members' work together.