# Red Giants: Chemical Abundance and Temperature

Anand Singh, Haoyuan Chen, Shaojie Zhang, Xiyan Chen

Abstract

A dataset provided by APOGEE which includes the information about 99705 red giants was used in the project. We proposed three questions about all red giants' chemical abundance and effective temperature and utilized different statistical tests and estimations to find the answers. Sample statistics collected from 99705 red giants were exploited to infer about the population parameters. A two-sample, two-sided hypothesis test was performed to investigate whether there's a difference between the mean effective temperature for stars with a negative abundance of magnesium and a positive abundance of magnesium. We found that there indeed exists a statistically significant difference. We speculated that magnesium might contribute to star's mass, and the mass would affect the effective temperature. Furthermore, red giants with different magnesium are at different stages of their evolutions, leading to different temperatures. We also wanted to find out whether Arcturus' effective temperature 4286K can represent the mean effective temperature. The bootstrapping method was applied and we could find the confidence intervals with different confidence levels to estimate the mean. Our result shows that 4286K is not contained within the confidence intervals, and the estimated mean is higher, roughly around 4630K. The population mean is very unlikely to be 4286K. We found a statistically difference and the result might help the astronomers to study about the different properties between Arcturus and other stars. In addition, we wonder whether the proportion of red giants that have a positive abundance of nitrogen is greater than 50%. We applied a one-sample, one-sided hypothesis test and concluded that we did not find strong evidence suggesting the proportion of red giants that have a positive abundance of nitrogen is greater than 50%. The result is not statistically significant but it provides insights into the formation and history of red giants.

## Introduction

The project focuses on the data of 99705 red giant stars observed by APOGEE, a stellar spectroscopic survey. We want to investigate the specific characteristics about chemical abundance and effective temperature of all the red giants in the universe. However, since we can't gain the data of all the red giants, we would use the observed data, which is the 99705 stellar spectra and related measurements from red giants collected by APOGEE, to infer the parameters of population which is all the red giant stars. We proposed three questions that intrigue us and took samples of different sizes from the APOGEE dataset to conduct different types of statistical tests and estimations, seeking for the answers to these questions. The study involves the use of hypothesis testing and bootstrapping. In addition, the whole coding process was done in R language. It's convenient to conduct the tests and visualize the results using R. To fully understand the report, the readers should be able to understand the statistical methods and R codes introduced in course STA130.

First, we want to investigate whether there is a difference in the mean effective temperature for red giants with the negative and positive abundance of magnesium. When the abundance of magnesium is positive, then the star contains more magnesium than our sun. When it's negative, then the star contains less magnesium than our sun. Magnesium is a type of element that can release huge amounts of heat when burned. It is one of the most abundant elements in the universe, and plays a crucial role in the formation and evolution of stars. It also has potential relationships with star's mass, age, and other compositions. We wonder whether the amount of magnesium would affect the effective temperature. This might provide us with information about red giant's formation and other potential characteristics. We decided to apply a two-sample hypothesis test with a permutation test to find out the answer.

Second, we want to test whether the mean effective temperature of red giants is equal to 4286K which is Arcturus' effective temperature. Arcturus is a well-known red giant; it's the brightest star in the constellation Boötes and one of the brightest stars we can directly see from Earth. We want to compare the effective temperature of Arcturus and other red giants in the dataset to see whether Arcturus can represent the mean effective temperature of red giants. Since Arcturus is a typical red giant star, its effective temperature might reflect the mean. The result might help astronomers to determine the average effective temperature of red giants. Since Arcturus is famous, it should be already studied and astronomers can thus use its information such as effective temperature and other properties to compare with other stars so that they might be able to find more relationships. For example, if Arcturus doesn't represent the mean, then what's the reason? Is it due to its brightness? Astronomers could conduct further research based on our result. To find the answer, we will apply bootstrapping to calculate the confidence interval and find the possible mean.

Finally, we want to explore whether the proportion of red giants that have a positive abundance of nitrogen is greater than 50%. A positive abundance means the red giant contains more nitrogen than our sun, while a negative abundance means the red giant contains less nitrogen than our sun. We are curious that whether most of the red giants contain more nitrogen than our sun. This can provide us with insights about the importance of nitrogen to the stars' formation and the distribution of elements among the stars. If half the red giants happen to have greater nitrogen than our sun, the scientists could investigate whether there's an underlying reason for this coincidence. A one-sample hypothesis test was used to test the null hypothesis that the proportion of stars that have a positive abundance of nitrogen is equal to 50%.

## Overview of data

Three continuous variables were used in the project: MG_H, N_H, and teff. MG_H and N_H represent the abundance of magnesium and nitrogen. teff represents the effective temperature. They all belong to APOGEE dataset. The units of MG_H and N_H are the base 10 logarithm units relative to the abundance of magnesium and nitrogen of our sun. The unit of teff is Kelvin. The rest of the report is divided into three sections. Each section corresponds to one question. The detailed processes about cleaning and wrangling the data will be introduced in three subsections of the report. Each section will include the content of analysis, results, data visualizations, discussion, and conclusion for one question. Please check both the final report and the rmd file "R codes" which includes all the reproducible codes and visualizations.

Question 1

Is the mean effective temperature for stars with negative abundance of magnesium the same as the mean temperature for stars with positive abundance of magnesium?

There are many different kinds of elements in the stars. And there is also a difference in the amount of elements among the stars. Magnesium is one of the most abundant elements in the universe, and plays a crucial role in the formation and evolution of stars. However, its impact on the effective temperature of stars is not yet fully understood. Especially when effective temperature is a new concept for us, we wanted to look deeper into whether the element Mg could have an effect on it.

Method:

The study uses a two-sample hypothesis test to compare the means of stars with negative and positive abundance of magnesium, and analyzes the data using box plot, histogram and other statistical methods. A two-sample hypothesis test is a statistical test used to compare the means, variances, or proportions of two samples. Also, we found that this question is falsifiable, which means that it is capable of being proven false. Furthermore, what we need to do is infer from the mean in the sample to that in the population and see if the results occurred by chance. We will conduct a two-sample hypothesis testing, which can help us to determine if there is a significant difference between two populations or samples.

Data:

To work with the useful data, in the first step is to do the pre-processing of the data. We need to extract two columns of data “MG_H” and “teff” from the database and use the cbind() function to merge them into a tibble named MG_teff, which represents the relative amount of magnesium on the surface of stars to the Sun and their corresponding temperature. Then we use the case_when function to name data with the relative amount of magnesium on the surface of stars to the Sun less than zero as “Negative” and those with a value greater than zero as “Positive”, which shows that there is more magnesium relative to our Sun and then use the select() function to store only the teff and Negative_or_Positive_MGH in MGTE.

Analysis and data visualization:

Now we have two groups. Notice that each group can be characterized by the mean of their effective temperature, we would assume Mpos, the mean teff of stars with positive MG_H is the same as Mneg, the mean teff of stars with negative MG_H. Therefore, we have our null and alternative hypotheses, where H0: Mpos - Mneg = 0 and H1: Mpos - Mneg != 0.

We then filter out the values that are not missing values to form a sample of size 28000 called sample_MGTE, which we use ggplot() to visualize it. Then we get this boxplot with the y-axis as the effective temperature and x-axis as the sign of MG_H, which we divided into a positive and a negative group. By comparing the boxes of these two groups, we can see that the range of effective temperature of positive

MG_H is much smaller, although there are some outliers around 4100 and 5250, its median is about 120 degrees lower than the one of negative MG_H. Plus, since both medians are located to the right of the box, both box plots are relatively skewed to the left. In terms of IQR, the distribution of MG_H data with positive values is more concentrated, so IQR is smaller than that with negative values, which are 250 and 500 respectively. By contrast, the effective temperature of the negative sign of MG_H has a wider spread of temperature data, up to about 5450 and down to about 4000.

Then we use the sample() function randomly sample the data and divide the data into two groups by positive and negative signs of MG_H by group_by() function. After that, we use summarize() function twice to find the mean values of the two groups and then make the difference, getting a value of -61.9, which means the mean of effective temperature for stars with more magnesium relative to our Sun is 61.9 smaller than that for stars with less magnesium relative to our Sun, which can confirm the previous conclusion reached by a boxplot.

Since this is a two-sample hypothesis test about finding the difference between means, we need a larger sample size and repeat the process of calculating different mean differences. We choose a permutation to help us randomly shuffle the data and create a distribution of the mean differences under the assumption that the null hypothesis is True, which allows us to calculate a p-value and make inferences about the population. Thus, we repeated the for loop 1000 times and put the mean of these 1000 sampling distributions together to form a large histogram. Then we can get a graph of all simulated test statistic data, which is the difference between mean effective temperatures of negative MG_H and positive MG_H. From this histogram, we can see that the simulated difference shows a quite symmetric normal distribution that is unimodal. The data is distributed from around -11 to 11, mainly concentrated around 0, which means that the mean is around 0 too. Since our alternative hypothesis is that the mean of them is not equal, so we should calculate the two-sided p-value, which is 0. We set the significance level 0.05, since $0 < 0.05$, we can reject the null hypothesis, which is that the difference between the mean effective temperatures of negative MG_H and positive MG_H is not equal to zero. This result is statistically significant, showing that the probability of obtaining a test statistic as extreme or more extreme than the observed result, assuming the null hypothesis is true, is extremely low. Therefore, the mean effective temperature for stars with the negative abundance of magnesium is not the same as the mean temperature for stars with the positive abundance of magnesium.

Discussion and Conclusion:

In other words, the abundance of magnesium in a star does have an impact on its effective temperature. By reviewing the information, we found that many factors can lead to changes in effective temperature, mass, age, composition, luminosity, etc. The most relevant ones to what we are exploring are mass and age. Therefore we can roughly infer that the different signs of magnesium abundance may have different masses, which play different degrees of influence on the effective temperature. Also, stars go through different stages of evolution as they age, so it is possible that the stars with negative magnesium abundance and those with positive magnesium abundance are at different stages of their evolution and this is contributing to the observed differences in their effective temperature.

However, the result does have limitations. We cannot exclude the existence of confounding variables, which means there may exist other factors that will also influence the effective temperature of stars. We can only make a rough inference based on the available information. Therefore, the observed differences in effective temperature between positive and negative MG_H groups may be due to other factors.

In conclusion, the abundance of magnesium in a star does impact its effective temperature and the differences in effective temperature between stars with negative and positive magnesium abundance may be related to their different masses and evolutionary stages.

Question 2

Is the mean effective temperature of the stars equal to 4286 Kelvin or not?

The second question is about the effective temperature, which is the surface temperature. One of the brightest red giants called Arcturus has an effective temperature of 4286K. We are trying to figure out if Arcturus' effective temperature can represent the mean effective temperature of all the red giants in the universe.

Method

We are using bootstrap to estimate the confidence interval to solve the problem. We need to test out the range of mean effective temperature value so that we will know whether the mean teff is around 4286 kelvin or not. Confidence interval is a range of estimates that the percentage of a set of values will be in. We are finding the mean value by not using the whole database so that using a small sample size but doing resamples with replacement get us many sample statistics to estimate the confidence interval.

Data

For data, we will extract the 'teff' out from the file that contains the data of 99705 red giants' effective temperature.

Analysis

First we will choose a sample from it randomly. We want to estimate the statistics with the sample, first set the seed to make sure that the output will be the same each time, then set a variable that randomly selects a sample with size 30000 from 99705 red giants. And we summarize the values in the variable and divide it by 30000. Then we got a tibble which shows the mean effective temperature is 4628 with size 30000. The first sample is for the initial estimation, and we still need more samples.

After estimating the sample, we need to resample with replacement and get sample statistics. Many test statistics can let us estimate a range of values, which is the confidence interval. We set seed again and have everything the same but this time the replacement equals true and got a single bootstrap sample with mean temperature of 4629 from 30000 red giants. Then resample them, we set a variable equal to 1000 to run the simulation and initialize a vector to store the means of each iteration. Next resample with replacement, repetition is the best way to estimate it because it will redo the code you put into the repetition and replay the code many times till the number you choose. Start the repetition with 1000 times, and sample with size 30000 and replacement equals true. Next compute the mean of the sampled values and store it in a vector. Finally, create a tibble with one column which shows the outcomes from bootstrap.

Confidence interval is a range of values with a certain level of confidence based on a sample of data. For the question, the confidence interval is containing a range of values of mean effective temperature, which if 4286 is not around the confidence level of range, then mean teff will unlikely be 4286 kelvin. We used quantities to compute the 5th and 95th percentiles of the values we resampled to check the confidence interval and we got 10% with 4626.264 and 90% with 4630.573. When trying other confidence intervals such as 99% intervals, we still get around 4624.625 to 4632.352.

## Results

As the confidence level we choose increases, the confidence interval doesn't change much. This means whatever we change, different samples with the same size, the interval is always around from 4600 to 4650, 4286 is very unlikely to be contained in the confidence interval. Therefore, the mean effective temperature of the red giants is very unlikely to be 4286 Kelvin.

## Data visualizations

Boxplot is one of the data visualizations we choose to describe the single sample mean effective temperature. Boxplot can directly see its median value, and we can see that for a random sample, the mean teff is around 4613 and its 1st quartile and 3rd quartile which its 25% to 75% value is around 4400 to 4800, the boxplot shows a left-skewed distribution and which means mean is less than the median, and its IQR is around 500. Also there is no outlier in the graph.

The second data visualization we used is Histogram, it can show the mean value directly, the variance, and the range of value. It shows the most shown mean value is around 4600-4625, and mean effective temperature range is around 4223 to 4234. The histogram is more or less a symmetry, in which the variance will be pretty small, so that illustrates the range will be pretty small. The histogram is showing many mean effective temperature values in a really small range and we can see that 4286 is not likely to be around the range.

## Discussion

The method we used is bootstrap resampling to estimate the sampling distribution of a statistic. The method is particularly useful when the underlying distribution of the population is huge or unknown, or when the sample size is small. The result isn’t significantly statistical based on no hypothesis and the result isn't 100% precise based on the limitations. Such as the sample might not be representative, the 30,000 red giants sample we choose is randomly selected from 'teff' data, we are not sure if the 30000 red giants sample is representative of the population of red giants or not. Also the distribution is normal with it is symmetry, unimodal and has small variance, if the sample distribution is a bimodal or having a big variance, then the result will not be as precise as normal.

## Conclusion

We first utilized a random sample of 30,000 stars, and bootstrap resampling was performed to estimate the sampling distribution of the mean effective temperature. The confidence intervals constructed from the bootstrapped sample statistics is way far from the value of 4286 Kelvin. Which indicates that 4286 is very unlikely to be the mean effective temperature of the red giants. Future work could be to explore the potential relationship between the mean effective temperature and other physical properties of stars, such as their age or composition.

## Question 3

Is the proportion of stars that have a positive abundance of Nitrogen greater than 50%?

Venturing deep into the cosmic realm, this study seeks to answer a burning question: Do more than 50% of red giants possess positive Nitrogen abundance? Hypothesis testing and simulations are employed as powerful tools to unravel this enigma. While a definitive conclusion remains elusive, the research provides valuable insights into the Nitrogen distribution in these celestial giants, laying a strong foundation for a more profound understanding of their chemical composition and origins.

Stars are the celestial bodies that harbor the secrets of their own evolution, and Nitrogen – a critical component – plays a significant role in shaping a star's life and behavior. The objective of this research is to delve into the complex world of Nitrogen abundance within red giants, aiming to discern the frequency of positive Nitrogen levels in these massive stars and shed light on their formation and history.

## Data & Methodology

Navigating the Vast Galactic Landscape: Using Nitrogen abundance data gathered from 99,705 red giants, this study examines a random sample of 35000 stars, seeking to determine the proportion of stars with positive Nitrogen abundance. A one-sided hypothesis test is employed, contrasting the null hypothesis, which posits that exactly 50% of red giants have positive Nitrogen abundance, with the alternative hypothesis which contends that more than 50% of red giants exhibit positive Nitrogen levels. Simulations are utilized to calculate the p-value, which serves as a means of evaluating the validity of the null hypothesis. The decision to use a one-sided hypothesis test stems from the research question's specific focus on determining whether the majority of red giants display positive Nitrogen abundance.

The histogram and bar plot - two potent visualization techniques - adeptly portray data distribution, shedding light on hypothesis testing.

Bar plot explained

Visualizing the dispersion of positive and negative Nitrogen abundance values in our simulated sample, the bar plot emerges as a powerful tool. By contrasting the proportion of stars with positive Nitrogen abundance to those with negative Nitrogen abundance, we gain valuable insights. In this instance, a relatively balanced distribution is evident. The bar plot excels in offering a distribution overview, enabling swift identification of any significant skew towards positive or negative Nitrogen abundance.

Histogram, demystified

Showcasing the distribution of simulated test statistics - the proportions of positive Nitrogen abundance across 1000 simulations - the histogram reveals the shape and spread of the distribution. Consequently, we pinpoint the observed test statistic (0.4986) within the distribution. Here, a fairly normal distribution emerges, centered around 0.5, with the observed test statistic nestled near the center.

Visualization analysis

The bar plot unveils the proportion of stars with positive Nitrogen abundance at nearly 50%, mirroring our null hypothesis. Reinforcing this observation, the histogram displays a distribution of simulated test statistics clustering around 0.5, and our observed test statistic lies within the distribution.

With a histogram-derived p-value of 0.697, surpassing the 0.05 significance level, we cannot reject the null hypothesis. This intimates that the proportion of stars with positive Nitrogen abundance isn't notably greater than 50%. In summation, the visualizations grant lucid comprehension of data distribution and hypothesis testing outcomes, indicating insufficient evidence to uphold the claim that over 50% of stars possess a positive Nitrogen abundance.

Revelations & Findings

Unveiling the Secrets of Stellar Nitrogen Content: The test statistic, which represents the proportion of the sample with positive Nitrogen abundance, measures 0.4986. Utilizing a thousand simulations under the assumption of the null hypothesis, a histogram is generated, depicting the distribution of test statistics. The resulting one-sided p-value is calculated to be 0.697. Interpreting the Results: Understanding the Implications of the P-value: Given that the p-value stands at 0.697, exceeding the 0.05 significance threshold, we are unable to reject the null hypothesis. This suggests that there is insufficient evidence to assert that more than 50% of red giants exhibit positive Nitrogen abundance.

## Wider Implications & Future Perspectives

This research endeavor contributes to our understanding of the Nitrogen content in red giants. Although it does not confirm that a majority of these stars possess positive Nitrogen abundance, further investigations employing larger samples or alternative methodologies could yield more conclusive findings. Examining other elements may also prove beneficial in enhancing our comprehension of the chemical makeup of stars. A more in-depth grasp of Nitrogen abundance in red giants has the potential to impact our understanding of stellar evolution and nucleosynthesis models, shaping the future of astrophysical research.

This study embarked on a captivating journey to uncover whether more than half of red giants contain positive Nitrogen abundance, employing hypothesis testing and simulations as the primary investigative tools. While it does not provide conclusive evidence, it serves as an essential stepping stone for further exploration. Future research may consider analyzing other elements or employing different approaches to gain a more comprehensive understanding of the complex and fascinating world of stars' chemical compositions and origins.

Our inquiry into red giants, despite not yielding statistically momentous outcomes, sheds light on Nitrogen prevalence within these celestial bodies. With a p-value of 0.697, exceeding the 0.05 threshold, we must accept the null hypothesis—thus, we cannot assert that a majority of red giants display heightened Nitrogen levels. However, the data's underlying structure offers valuable insights, laying the groundwork for continued investigation. The techniques employed—hypothesis tests and simulations—proved potent in examining and deciphering data. By facilitating Nitrogen distribution analysis in red giants, they also enabled p-value calculation, essential for null hypothesis validity assessment. Visual aids, such as bar plots and histograms, further simplified data comprehension, fostering an intuitive grasp of the findings.

While not statistically significant, our discoveries supply a vital backdrop for ensuing research. Alternate methodologies, expanded samples, or investigations into other elements' abundance might yield a more exhaustive understanding of red giants' chemical makeup. These endeavors could profoundly impact stellar evolution comprehension and nucleosynthesis models, molding the course of astrophysical inquiry.

Within the grander scheme, our study paves the way for future examinations into stars' chemical compositions and origins. Although not statistically significant, the findings enrich our knowledge of Nitrogen distribution in red giants, opening avenues for further exploration. Ultimately, this research accentuates the significance of relentless investigation and discovery in astrophysics, pushing us to unravel the cosmos' enigmas.

## Citation

### MLA

“Effective Temperature.” Wikipedia, Wikimedia Foundation, 22 Feb. 2023, https://en.wikipedia.org/wiki/Effective_temperature.

“Two-Sample Hypothesis Testing.” Wikipedia, Wikimedia Foundation, 27 Apr. 2022, https://en.wikipedia.org/wiki/Two-sample_hypothesis_testing.

Information@eso.org. “Red Giant.” ESA/Hubble | ESA/Hubble, https://esahubble.org/wordbank/red-giant/.

“Arcturus.” Wikipedia, Wikimedia Foundation, 18 Mar. 2023, https://en.wikipedia.org/wiki/Arcturus.

## Appendix

Several changes to the sample sizes were made after submitting the progress report. Our original thought was to view 99705 red giants as our population and we would conduct various statistical methods to answer the questions we proposed. The original sample sizes we took were extremely small: around 100. However, we realized that we already gained the data of 99705 red giant stars. Then there’s no need to apply these statistical methods. We then noticed that these 99705 red giants are just a subset of the population, all the red giant stars, and we are inferring the population parameters using the statistics. Therefore, we assume the data of 99705 red giants is representative and we want to take larger samples from it. Therefore, we increased the sample size to approximately 30000 since this will ensure we don’t take extreme samples and increase accuracy. The result for questions 2 and 3 do not have significant changes. However, for question 1, when the sample size was around 100, we failed to reject the null hypothesis, but now we reject the null hypothesis. As the sample size increased, we obtained more precise results and avoided the possible type 2 error. We then fixed the related R codes and obtained new data visualizations.

## Individual Contribution:

In the course of this project, I undertook a multitude of tasks, chiefly delving into research, dissecting data, and crafting the report for the third question. My contributions spanned several key areas:

Probing and Scrutiny:

- Engaged in amassing data on Nitrogen levels within a staggering 99,705 red giants, subsequently narrowing it down to a random assortment of 35,000 celestial bodies for our examination.
- Teamed up with colleagues to devise and carry out a one-sided hypothesis test, pinpoint the null and alternative hypotheses, and conduct simulations to compute the p-value.
- Took the reins in dissecting the bar plot and histogram visual aids, enabling a clearer grasp of data distribution patterns and the fruits of our hypothesis testing.

Penning the Report:

- Crafted the report's introduction, data & methodology, visualization scrutiny, insights & discoveries, and broader implications & prospective directions sections.
- Ensured a seamless narrative flow, effectively conveying the research inquiry, methodology, results, and ramifications.

Team Synergy:

- Fostered an environment of collaboration, discussing research inquiries, exchanging thoughts, and fine-tuning our investigative methods.
- Offered constructive criticism on disparate sections of the report, partaking in the editing process to guarantee a polished, cohesive final product.

Through this endeavor, I honed vital skills in data dissection, hypothesis testing, and simulation methodologies. I delved deeper into the nuances of Nitrogen abundance in red giants, appreciating the profound impact such inquiries have on the field of astrophysics. Additionally, I refined my aptitude for conveying intricate ideas via visual aids and succinct writing.

My most daunting challenge lay in interpreting the data such that it was both accurate and accessible to a diverse audience. To surmount this, I turned to visualizations and employed lucid language to express our findings. Moreover, I consulted teammates to ensure my interpretations aligned with our research goals and maintained accuracy.