Jay Frothingham

Project 2

Background

Exoplanets are planets beyond our own solar system which orbit stars other than our own Sun. We are able to detect exoplanets through a variety of methods, including taking direct images, measuring changes in stellar brightness as planets orbit, and measuring shifts in spectra as planets orbit.

The exoplanet database is a collection of information about exoplanets. It contains properties of exoplanets as well as their host stars ranging from the quantitative to qualitative data. It includes orbital periods and radii, masses, densities, coordinates, discovery methods, discovery years, number of confirmed detections, and discovery facility, to name a few. The exoplanet database is hosted and maintained by NASA, though it contains measurements from a variety of exoplanet surveys.

The main data explored in this project are the stellar mass and number of planets that make up a planetary system. Stellar mass is measured in terms of solar mass, and can be determined by measurements of stellar luminosity. The number of planets in a system can be determined in different ways depending on the method by which the system and planets have been discovered.

This report will investigate whether systems with more massive host stars are more likely to have more exoplanets. By Newton's Law of Universal Gravitation, force is proportional to mass, so more massive stars exert a greater gravitational force.

Newton's Law of Universal Gravitation applied to a host star and exoplanet

$F = G (Mm / r^2)$ where *F is the force between the host star and exoplanet*
M is the mass of the host star
m is the mass of the exoplanet
r is the distance between the host star and exoplanet

With this in mind, I was curious whether more massive stars could be more successful at keeping multiple planets in a stable orbit.

Of the systems in the exoplanet database, the ones which have data available for both stellar mass and number of planets in the system are mainly those discovered by the transit method and by radial velocity. This is a potential source of bias in the dataset. It's possible that the detection methods favor particular types of systems, so some system types may be overrepresented or underrepresented in the database compared to their actual physical occurrences.

Procedure

To complete this investigation, I determined that the maximum number of planets per system in the database was eight planets. I split the data into populations of one, two, three, four, five, six, seven, and eight planets so that I could plot and analyze the different types of systems. I further split each of these populations into subpopulations by discovery method so that I could visualize and compare the different distributions of the discovery methods for each population. I extracted the stellar masses for each population and subpopulation.

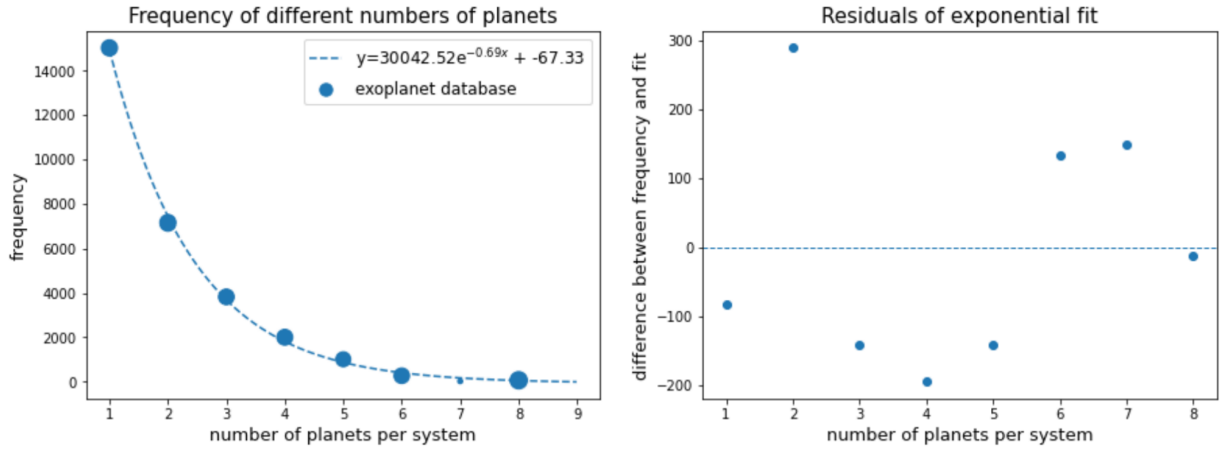
I calculated the frequency of different numbers of planets in the dataset, and calculated the mean stellar mass for each population of different numbers of planets. I also calculated "weights" based on the amount of stellar mass data in each population.

The code used to manipulate the data and generate plots was written in Python. It utilized the Python libraries `pandas` to load and manipulate the dataset, `numpy` and `scipy` for general numeric and statistical operations, and `matplotlib` for plotting, as well as a few custom-written functions to find the frequencies of values in a dataset and define an exponential decay function for curve fitting.

Discussion and analysis

To get an idea of the distribution of systems with different numbers of planets in the database, I plotted the frequency of each population.

Figure 1. Frequency of systems with different numbers of planets.



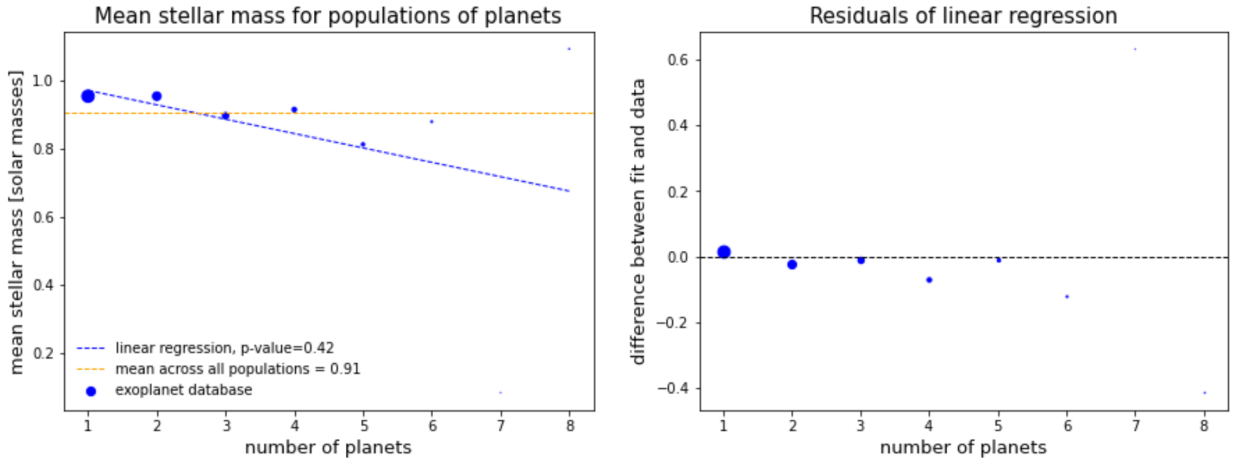
Markersize indicates mean stellar mass for systems with different numbers of planets.
The figure shows the frequency of systems with different numbers of planets, and an exponential fit to describe the relationship between frequency and number of planets. The residuals of the exponential fit are also plotted.

Most of the systems in the exoplanet database have only one planet. The frequency of systems with different numbers seems to drop off exponentially. Systems with more planets occur less frequently in the exoplanet database.

Initially, the residuals of the exponential fit appear poor. A few more of the residuals are negative than positive, indicating that the model is a slight underestimate compared to the data, though there does not appear to be any kind of pattern to the residuals. Very few of them are visually close to zero, which would indicate a perfect fit. However, the scale of the plotted frequencies is on the order of 1e3 to 1e4, and the residuals are on a scale of 1e2. Considering orders of magnitude, the residuals actually do support the exponential model as a good fit.

The plot also includes markersizes scaled to mean stellar mass, as a way to quickly compare across populations. Based on Figure 1, the mean stellar mass appears visually similar across populations. The population of seven-planet systems has a noticeably smaller mean stellar mass, but excluding that data point, it is difficult to visually tell from the markersize which populations have larger or smaller mean stellar masses. So I plotted the mean stellar mass against the number of planets for different populations.

Figure 2. Mean stellar mass for systems with different numbers of planets.



Markersize indicates how many data points were used to calculate the mean stellar mass. The figure shows the mean stellar mass for each population of systems with different numbers of planets and a linear fit to the relationship between number of planets and mean stellar mass. Residuals for the linear fit are also plotted.

Visually, the linear fit has a negative slope, suggesting that systems with greater numbers of planets have lower mean stellar masses. It's likely that the outlier mean stellar mass of systems with seven planets is impacting the linear fit. Because there are fewer systems in the exoplanet database with high numbers of planets, there are fewer data points available to calculate the mean stellar masses of those populations. This means that the mean stellar masses are less reliable for populations of systems with higher numbers of planets. To indicate that populations

with fewer data points are less reliable, the mean stellar masses are plotted with markersizes proportional to the amount of data used to calculate them.

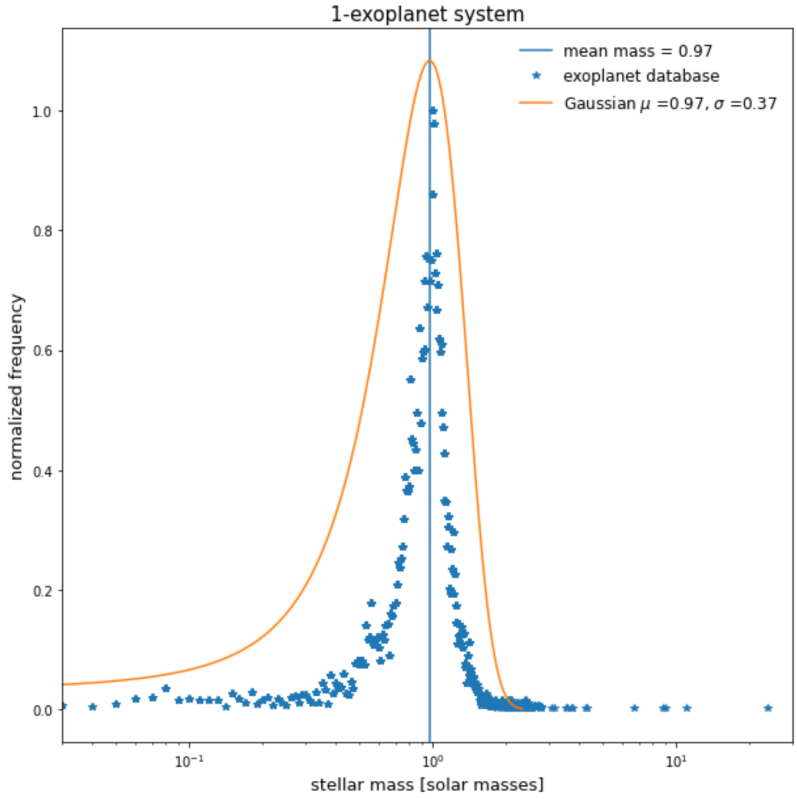
The residuals of the linear regression seem to indicate a good fit. With the exception of populations with significantly fewer data points (six, seven, and eight planets), the residuals are all very close to zero. Most of the residuals are less than zero, indicating that the linear regression is a slight underestimate. There is no noticeable pattern in the residuals.

The Python function used to generate the linear regression also calculates a p-value, assuming a null hypothesis of a horizontal line. The alternative hypothesis would be that the data follows a line with non-zero slope. The calculated p-value for the linear regression was 0.42. A p-value of zero indicates agreement with the alternate hypothesis, while larger p-values indicate agreement with the null hypothesis. A significance level of 0.05 is fairly standard, where p-values greater than 0.05 indicate agreement with the null hypothesis and less than 0.05 indicate agreement with the alternative hypothesis. In this case, the p-value is substantially greater than 0.05. The p-value strongly suggests that the mean stellar mass of a system remains constant rather than increasing or decreasing with the number of planets in the system.

The data are also plotted with a horizontal line representing the mean stellar mass across all populations, calculated to be 0.91 solar masses. As the systems with higher numbers of planets do not contain as many data points, their impact on the calculation is less likely to seriously skew the average. The mean stellar masses of systems with fewer planets (one or two planets) are higher than this average. The mean stellar masses of systems with three or four planets are very close to this average. The mean stellar masses of systems with more planets (five or six planets) are lower than this average. However, the mean stellar masses of all systems with one to six planets were within 0.1 solar masses of the average.

As many of the populations seemed to have similar mean stellar masses, I wondered whether they also had similar distributions, and whether certain detection methods made up different parts of the distributions.

Figure 3. Distribution of stellar masses for systems with one planet.

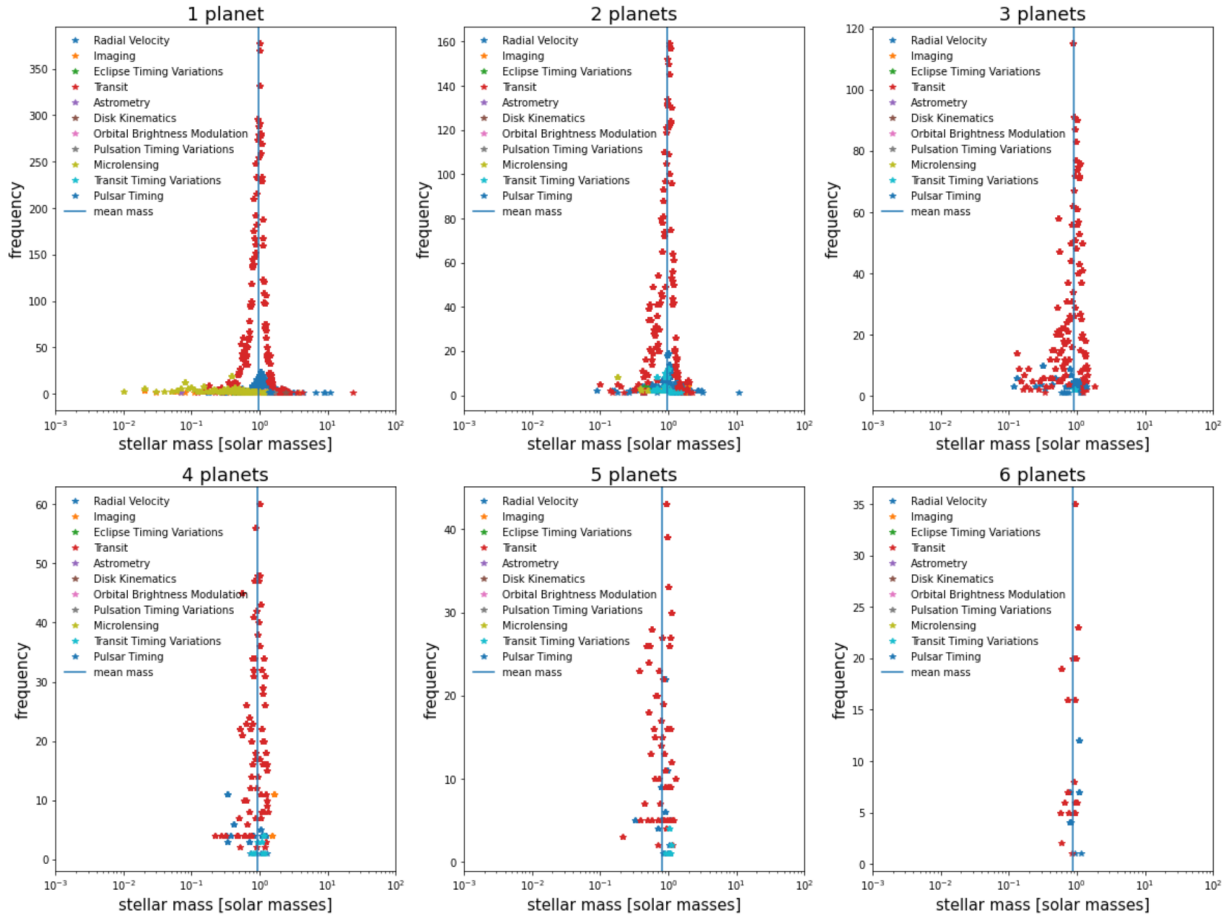


The figure shows the distribution of stellar masses for systems in the exoplanet database with one planet. The data have been normalized to one, and are plotted with a Gaussian distribution centered at the mean stellar mass of one-planet systems in the database.

Python functions were used to estimate the mean and standard deviation of the data. A different Python function was used to generate a Gaussian distribution with the same mean and standard deviation. By visual inspection, the data qualitatively is similar to the plotted Gaussian distribution. The Gaussian distribution is not a good fit—very few of the data points lie on the plotted Gaussian—but its general shape and curvature match.

If the distribution of stellar masses around the mean for each population has approximately the same mean and standard deviation, that would suggest that stellar mass and number of planets are unrelated. So I plotted distributions for each population of systems with different numbers of planets. I also indicated the method of discovery for each data point in the distributions, to see whether there were any patterns that could impact the distributions for the populations.

Figure 4. Distributions of stellar mass color-coded by detection method.



The figure shows distributions of stellar mass for systems with one to six planets. The data in each distribution is color-coded by discovery method.

By visual inspection, it is possible that systems with more than one planet also have a Gaussian distribution of stellar mass. The distributions for systems with greater numbers of planets do display stellar masses approximately centered around similar means, with very high frequencies for stellar masses near the mean and lower frequencies for stellar masses further from the mean. However, the number of data points decreases dramatically for systems with more than one planet. For systems with more than one or two planets, there is not enough data to convincingly show a Gaussian distribution, or really any formal distribution. Because the distributions are unclear, the figure does not strongly support a lack of relationship between stellar mass and number of planets. Nor does it support a relationship. More data for systems with multiple planets would be needed to make a claim either way.

In terms of discovery method, the overwhelming majority of systems in the plotted distributions were discovered using the transit method. The transit method is responsible for detections of systems with stellar masses both close to and to either side of the mean, and with the greatest frequencies. Radial velocity and transit timing variation are responsible for the next most detections, with much lower frequencies, of stellar masses spread around the mean. Systems

with one planet display a lot of detections through microlensing of low-frequency stellar masses lower than the mean. Systems with more than one to three planets had few enough data points that splitting the population by discovery method didn't yield particularly meaningful information.

In conclusion, this project did not find evidence to support the idea that exoplanet systems with more massive host stars are more likely to have multiple planets. Plotting mean masses against numbers of planets and fitting a possible linear relationship resulted in a p-value that supported the null hypothesis stating that the mean mass is a constant value unrelated to the number of planets. And plotting the distributions of stellar masses for different populations was uninformative. Due to a lack of data for systems with higher numbers of planets, the distributions could not be rigorously compared. As exoplanet detection methods develop, it is possible that the necessary data will be collected, and this question could be explored more in the future.