Examining Biases in Exoplanets Detections
How Typical is our Solar System?

Tal Sharoni and Justin Lipper
McGill University Departement of Physics
(Dated: April 19, 2024)

Recent decades have witnessed remarkable strides in exoplanet research, discovering diverse worlds far beyond our own solar system, each with different properties and features. Leveraging innovative detection techniques as well as state of the art telescopes like the JWST, we have now discovered around 5000 exoplanets in total, as per NASA's latest update, as well as thousands more potential exoplanets candidates.

Five primary detection methods, radial velocity, transit, direct imaging, gravitational microlensing, and astrometry have been used to identify exoplanets candidates, however each of those methods introduce their own unique biases, which can make the types of discovered exoplanets unrepresentative of reality. In this project, we aimed to correct the biases of two of those detection methods - radial velocity and transit methods and using the planets of our solar system as a representative sample (since those planets were not subject to any sort of detection bias) we used Bayesian model selection to see if our solar system is in fact a representative sample, we did this Bayesian analysis on both all the planets, and also just the inner planets.

We have concluded that with our corrections, the corrected models generally provide a better representation of reality compared to the uncorrected models, particularly for orbital periods and semi-major axes. However, discrepancies in the correction for planetary radii suggest the need for further refinement in our approach. If we only consider inner planets however, the planetary radii parameter doesn't appear better or worse (while the orbital periods and semi-major axes still remain better post-correction), and the stellar radii are trivially unaffected by our corrections. We also concluded that to obtain more representative samples of exoplanets orbital periods and semi-major axes, exoplanets as a whole would need to be detected for much longer than we currently have.

I. INTRODUCTION

The past few decades have been abound with advancements in exoplanet science, giving astrophysicists a glimpse into the types of worlds that exist not only in our own solar system, but throughout our stellar neighborhood. By making use of a few clever detection techniques and state of the art telescopes such as the Kepler space telescope and James Webb space telescope, scientists have been working to gain a better understanding what types of planets exist in the Universe and how common each type is. Despite these advancements however, observational bias presents a major hurdle standing in the way of a statistically sound understanding of the population of exoplanets.

The five methods that have been used to detect exoplanets are the radial velocity method, transit method, direct imaging, gravitational microlensing, and astrometry. Each of these methods takes advantage of a different imprint that an orbiting planet leaves on light from the star system of interest that makes its way to Earth, meaning each method has its own strengths and weaknesses. The radial velocity and transit methods comprise the vast majority of exoplanet detections, since the latter three require very specific system configurations and technology that has not yet been developed to its full potential. Being by far the most utilized methods for exoplanet detection, understanding and correcting for the observational biases inherent to these methods will greatly improve our understanding of the true population of exoplanets.[1]

A. Radial Velocity Method

The radial velocity method makes use of the fact that an orbiting planet exerts a gravitational tug on its host star, causing it to change velocity slightly over the course of the orbital period. In fact, the star and the planet both orbit their common center of mass. If a component of this velocity change is colinear with the line of sight to Earth, this change in velocity can be detected as Doppler shift in the stellar spectrum. The orbital velocity of the star around center of mass (v_*) is given by the equation,

$$v_* = \frac{M_p}{M_* + M_p} v_k, \tag{1}$$

where M_p is the mass of the planet, M_* is the mass of the star, and v_k is the Kepler Velocity defined as

$$v_k = \sqrt{\frac{G(M_* + M_p)}{a}}, \tag{2}$$

where G is the gravitational constant and a is the semi-major axis of the planet's orbit.[2]

B. Transit Method

When a star system is aligned such that a planet passes directly between the star and Earth, a small amount of starlight heading towards Earth is blocked by the planet. This transit results in a periodic dip in the observed light

curve, which can be detected by measuring the flux of the host star as a function of time. To first order, the proportion of starlight blocked by a transiting planet is given by the equation,

$$\frac{\Delta F}{F} = \frac{R_p^2}{R_*^2}, \tag{3}$$

where ΔF is the difference in stellar flux induced by the transit, F is the flux of the star, R_p is the radius of the planet, and R_* is the radius of the star.[3]

In order to observe a transit, the planet must pass over a portion of the stellar disk from the point of view of Earth. The impact parameter (b) is a quantity defined to measure the minimum (over the orbital period) apparent offset between the center of the stellar disk and the center of the planetary disk, as seen from Earth. An impact parameter of zero implies that the planet passes in front of the center of the stellar disk while an impact parameter greater than one means that the planet misses the stellar disk entirely, meaning no transit is observed. Assuming a circular orbit, the impact parameter is defined by the equation,

$$b = \frac{a * cos(i)}{R_*}, \tag{4}$$

where a is the semi-major axis of the planet's orbit, i is the inclination of the orbital plane with respect to Earth, and R_* is the radius of the star.[3]

C. Goals

This project sought to correct for the biases introduced by the radial velocity and transit methods in order to gain a better understanding of the true population of exoplanets. From a catalogue of exoplanets found with the radial velocity or transit method [4], distributions of a few important parameters relating to the planets were constructed, and then corrected according to the observational biases associated with each method. To test how successful the corrections were, our Solar System was chosen as a representative sample of exoplanets, since it is the only set of planets whose discoveries are not subject to exoplanet detection biases. We determined whether or not our Solar System seemed more typical according to the corrected model, which according to Bayesian statistics would indicate the new model is likely to better describes reality.

II. METHODS

A. Obtaining data

The data was obtained from the NASA Exoplanet Archive [4], which consists of 5,609 discovered exoplanets and their known parameters. The data was filtered to remove planets not discovered with the radial velocity or transit method or missing any of the parameters used in the analysis, which are planetary radius, stellar radius, orbital period, semi-major axis, planetary mass, and stellar mass. The resulting list included 1,089 exoplanets.

B. Fitting Functional Forms for Parameters

With the exoplanets data obtained from NASA Exoplanet Archive, we first plotted four different histograms, one for each parameter and used a distribution function to model those parameter distribution. It's worth noting that the data of the planets was filtered, firstly if any on the fields of interest was missing (since we needed planets with all the fields for the comparison) and also planets with an orbital period of more than 600 days and a semi major axis bigger than 1 (earth's semi major axis) this was done because there were far too few planets with those parameters higher than what we specified, except outliers with extremes values and with them included it messed up our histogram bins since there would just be one big bin near zero and really smalls (almost invisible) bins for any higher values.

We estimated the model of the distributions of those parameters by using a bimodal Gaussian distribution for the radii

$$\begin{aligned}\text{bimodal}(x, \mu_1, \sigma_1, \mu_2, \sigma_2, w_1) = \\ \frac{w_1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) \\ + \frac{(1-w_1)}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right)\end{aligned} \tag{5}$$

where our function arguments represent:

μ_1, μ_2 : Means of the two Gaussian components
σ_1, σ_2 : Standard deviations of the two Gaussian components
w_1 : Weight of the first Gaussian bump in the distribution

and a lognormal distribution for the other three parameters

$$\text{lognormal}(x, m, \sigma) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln(x) - \ln(m))^2}{2\sigma^2}\right)$$

where:

m : Median of the log-normal distribution
σ : Standard deviation of the log-normal distribution

Refer to the Appendix section A for the fitted distributions pre-corrected and post-corrected distributions.

C. Estimating Biases

Two types of detection biases were considered for corrections, biases from the intrinsic properties of the planet (planet property biases) and biases from the orbit of the planet (orbital biases). Planet property biases relate to the planetary radius and planetary mass while orbital biases relate to the semi-major axis and orbital period. It should be noted that correlation between these two types of biases was not considered as a simplifying assumption.

1. Planetary Properties Bias for Radial Velocity Method

It is assumed that likelihood of detecting a planet is proportional to the orbital velocity of the star about the center of mass, since larger orbital velocities induce larger and therefore more detectable Doppler shifts. Using Equation 1, we can see that v_* scales as

$$v_* \alpha \frac{M_p}{M_* + M_p} * \sqrt{M_* + M_p} = \frac{M_p}{\sqrt{M_* + M_p}}, \quad (6)$$

where the semi-major axis a was taken to be constant as it is an orbital property, so we express the planetary properties bias for the radial velocity method ($bias_{p,rv}$) as

$$bias_{p,rv} = \frac{M_p}{\sqrt{M_* + M_p}}. \quad (7)$$

2. Planetary Properties Bias for Transit Method

The detectability of a transiting planet is strongly dependent on the ratio between planetary radius and stellar radius, as shown in Equation 3. This dependence is quite intuitive, as this equation represents the proportion of the flux-emitting stellar disk that is blocked by the planet. In the publication *Observational Biases for Transiting Planets* [5], the planetary properties bias of a planet detected with the transit method ($bias_{p,tr}$) estimated to be

$$bias_{p,tr} = (\frac{R_p}{R_*})^{\frac{5}{2}}. \quad (8)$$

3. Orbital Properties Bias for Radial Velocity Method

From Equations 1 and 2, we can see that the orbital velocity of the star about the center of mass scales as,

$$v_* \alpha \sqrt{\frac{1}{a}}, \quad (9)$$

where only quantities for orbital properties are considered as non-constants. The orbital properties bias for the radial velocity method ($bias_{o,rv}$) is taken to be,

$$bias_{o,rv} = \sqrt{\frac{1}{a}}. \quad (10)$$

4. Orbital Properties Bias for Transit Method

Two sources of bias relating to the orbital properties were considered for the transit method. Firstly, the frequency of a transit (f) is inversely proportional to the orbital period of a planet. A small orbital frequency is very helpful for exoplanet detection, both by increasing the chances that a transit will occur while a telescope is surveying the host star and by making repeat observations of transits possible. Repeat observations of transits can be stacked on top of each other to increase the strength of the detection signal and are also necessary to confirm that a dip in the light curve is indeed due to a transiting planet. We estimated this bias favoring the detection of planets with small orbital periods ($bias_T$) as

$$bias_T = \frac{1}{T}, \quad (11)$$

where T is the orbital period of the planet. The second source of bias related to the impact parameter (b), introduced in Equation 4. In order to observe a transit, b must be less than one. It is evident from Equation 4 that range of acceptable values of $cos(i)$ such that this condition is satisfied is dependent on the ratio $\frac{a}{R_*}$, with smaller values of the ratio allowing the transit to be visible for a wider range of $cos(i)$, and thus more likely to be observed. We estimated the bias relating to the impact parameter ($bias_b$) to be the inverse of this ratio,

$$bias_b = \frac{R_*}{a}. \quad (12)$$

To obtain the overall orbital properties bias for the transit method ($bias_{o,tr}$), we take the product of $bias_T$ and $bias_b$ to obtain,

$$bias_{o,tr} = \frac{1}{T} * \frac{R_*}{a}. \quad (13)$$

D. Applying Corrections for Bias

Once each type of bias is computed, new weighted histograms were constructed for each parameter, with each planet in the dataset weighted according to its detection method by a correction term for biases relating to both the planetary property parameters and relating the orbital parameters. The correction term for planets detected through the orbital velocity method ($corr_{ov}$) was computed as,

$$corr_{ov} = \frac{1}{bias_{p,ov}} * \frac{1}{bias_{o,ov}}. \quad (14)$$

The correction term for planets detected through the transit method ($corr_{tr}$) was computed as,

$$corr_{tr} = \frac{1}{bias_{p,tr}} * \frac{1}{bias_{o,tr}}. \quad (15)$$

The weighted histograms were fit with the same functional forms as in Section II B to obtain a new set of probability distribution functions for the four parameters.

E. Obtaining Bayesian Likelihood and Posterior for our Solar System

In its most abstract form, Bayes Theorem states that,

$$\text{posterior } \alpha \text{ likelihood} * \text{prior}, \tag{16}$$

where

$$\text{posterior} = P(\text{model describes reality}|\text{observed data}), \tag{17}$$

$$\text{likelihood} = P(\text{observed data}|\text{model describes reality}), \tag{18}$$

and

$$\text{prior} = P(\text{model describes reality}). \tag{19}$$

We apply the continuous form of Bayes Theorem, where the posterior, likelihood, and prior represent continuous probability distributions of some parameter rather than discrete probabilities. A separate analysis is done for each of four parameters (planetary radius, stellar radius, semi-major axis, and orbital period).

Bayes Theorem is applied as follows: our models are the fitted probability density functions for the parameter of interest using first the uncorrected and then corrected data from the exoplanet catalogue, and our observed data consists of the values of the parameter of interest for each of the planets in our Solar System[6][7]. We assume that we have no prior knowledge of the distribution of the parameter of interest among planets of the Universe, implying a uniform prior.
To obtain a likelihood, or in other words to determine where our Solar System would lie on a probability density function for the parameter of interest for systems with eight planets, we assume that each planet represents an independent data point and compute the likelihood as follows:

$$p(\text{Solar System}|\text{model}) = p(\text{Mercury}|\text{model}) \cdot p(\text{Venus}|\text{model})..., \tag{20}$$

where *model* refers to the probability density function for the parameter of interest. Since we have assumed the prior to be uniform, the posterior is simply equal to the likelihood implying

$$p(\text{model 1}|\text{Solar System}) \; \alpha \; p(\text{Solar System}|\text{model 1}), \tag{21}$$

and

$$p(\text{model 2}|\text{Solar System}) \; \alpha \; p(\text{Solar System}|\text{model 2}), \tag{22}$$

where model 1 represents the uncorrected probability density function and model 2 represents the corrected probability density function.

F. Comparing the Models Using Bayes Factor

In order to compare the correctness of our two models (before corrections and after corrections) for each parameter of interest, we make use of Bayes Factor, defined as the ratio of the posteriors of the models, or

$$\frac{P(\text{model 2}|\text{Solar System})}{P(\text{model 1}|\text{Solar System})}. \tag{23}$$

A Bayes Factor much greater than one implies that the corrected model is a much more accurate representation of the true distribution of the parameter of interest than the uncorrected model. A Bayes Factor much closer to zero than one implies that the corrected model is a much less accurate representation of the true distribution than the uncorrected model. A Bayes Factor on the order of one implies that neither model is strongly preferred by the data over the other.

G. Considering only Inner Planets

Upon inspection of the data by eye, it became clear that the overwhelming majority of exoplanets discovered had orbital periods and semi-major axes much smaller than those for the outer planets of our Solar System, which makes sense given the detection biases discussed in Section II C. The bias against planets with orbital periods and semi-major axes similar to the outer planets is so extreme in fact, that even our corrections would not be likely to account for their existence since a complete lack of data cannot be corrected by any scaling factor. To account for this fact, we did the Bayesian analysis discussed in Sections II E - II F again only considering the inner planets of our Solar System as our data, Mercury, Venus, Earth, and Mars.

H. Plotting the Solar system planets against the data/correction

To look at the bigger picture, we decided filter out our data so that it only includes G-type stars like our sun, we set a stellar radius range of 0.9-1.1 suns, and plotted the density of the 3 parameters against the planets in our solar system. This was done in order to visually see how prevalent planets like those in our solar system actually are. This was done for both the uncorrected data and our corrected projections in order to compare.

III. RESULTS

A. Planetary Radii

TABLE I: Bayesian Model Selection Results for Planets Radii

	All Planets	Inner Planets
Model 1 Posterior	$1.1938 \cdot 10^{-10}$	$2.0221 \cdot 10^{-4}$
Model 2 Posterior	$5.6017 \cdot 10^{-21}$	$2.9875 \cdot 10^{-4}$
Bayes Factor	$4.6924 \cdot 10^{-11}$	$1.4774 \cdot 10^{0}$

B. Stellar Radii

TABLE II: Bayesian Model Selection Results for Stellar Radii

	All Planets	Inner Planets
Model 1 Posterior	$1.2164 \cdot 10^{0}$	$1.2164 \cdot 10^{0}$
Model 2 Posterior	$8.6518 \cdot 10^{-1}$	$8.6518 \cdot 10^{-1}$
Bayes Factor	$7.1124 \cdot 10^{-1}$	$7.1124 \cdot 10^{-1}$

C. Orbital Periods

TABLE III: Bayesian Model Selection Results for Orbital Periods

	All Planets	Inner Planets
Model 1 Posterior	$2.4855 \cdot 10^{-80}$	$1.0210 \cdot 10^{-20}$
Model 2 Posterior	$4.8518 \cdot 10^{-54}$	$3.7473 \cdot 10^{-15}$
Bayes Factor	$1.9521 \cdot 10^{26}$	$3.6704 \cdot 10^{5}$

D. Semi-major Axes

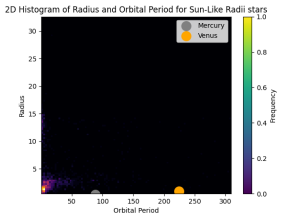
TABLE IV: Bayesian Model Selection Results for Semi Major Axes

	All Planets	Inner Planets
Model 1 Posterior	$4.1515 \cdot 10^{-53}$	$1.5932 \cdot 10^{-9}$
Model 2 Posterior	$1.9806 \cdot 10^{-25}$	$3.8742 \cdot 10^{-4}$
Bayes Factor	$4.7709 \cdot 10^{27}$	$2.4316 \cdot 10^{5}$

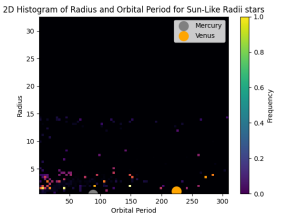
E. 2D Histograms against our solar system planets

IV. DISCUSSION

When considering all planets in our Solar System as data, our results generally suggest that the corrected model is a better representation of reality than the uncorrected model, with the exception of one parameter. For orbital periods, the calculated Bayes Factor of $1.9521 \cdot 10^{26}$ is larger than one by many orders of magnitude, implying that the corrected model is significantly more likely to represent reality than the uncorrected model. For semi-major axes, the Bayes factor was calculated to be $4.7709 \cdot 10^{27}$, which also indicates that the corrected model is significantly more likely to represent reality. These results are expected and suggest that our corrections provide a better model

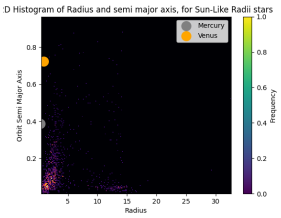


(a) 2D histogram of distribution density of orbital period vs radius (uncorrected)

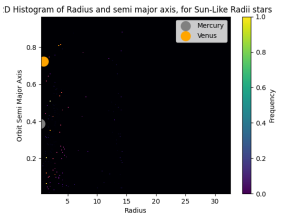


(b) 2D histogram of distribution density of orbital period vs radius (corrected)

FIG. 1: Comparison of 2D histograms for orbital period vs radius

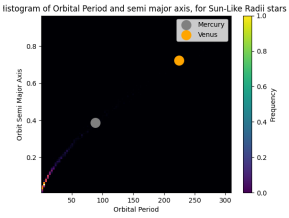


(a) 2D histogram of distribution density of radius vs semi major axis (uncorrected)

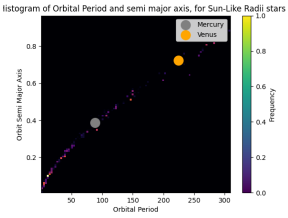


(b) 2D histogram of distribution density of radius vs semi major axis (corrected)

FIG. 2: Comparison of 2D histograms for radius vs semi major axis



(a) 2D histogram of distribution density of orbital period vs semi major axis (uncorrected)



(b) 2D histogram of distribution density of orbital period vs semi major axis (corrected)

FIG. 3: Comparison of 2D histograms for orbital period vs semi major axis

of the true distributions of exoplanet orbital periods and semi-major axes than the uncorrected data provide. However, it would be incorrect to conclude that our corrected model is the best model possible, since our analysis only compared two models. A hypothetical third model could fit the data better than both of the analysed models.

For planetary radii, the calculated Bayes Factor of $4.6924 \cdot 10^{-11}$ suggests that the corrected model is significantly less likely to represent reality than the uncorrected model. This result is unexpected and indicates an inaccuracy in the correction applied for the biases dependent on planetary radii. A likely explanation for this discrepency is our assumption that the planetary properties parameters and orbital parameters are independent. Our correction assumes that large planets should be easier to find since they block more light during transits and result in a more significant change in the host star's radial velocity. To correct for this, the correction significantly scales down the expected number of large planets in the Universe.

However, we did not consider that large planets tend to form beyond the ice line, farther from their stars. As previously discussed, planets with large semi-major axes and orbital periods are very difficult to detect. Taking this correlation between planetary size and distance from the star into account would have allowed for consideration of this source of bias, increasing the predicted number of large planets, and making our Solar System seem anomalous. This would have resulted in a higher posterior for the corrected model and a higher Bayes Factor.

For stellar radii, the calculated Bayes Factor is $7.1124 \cdot 10^{-1}$. This value is on the order of one and suggests that there is no strong preference between the models. This is expected since the corrections were meant to address biases in planet detection, not star detection, not star detection, so applying the corrections should not significantly affect the distribution of stellar properties.

When considering only the planets in the inner Solar System, the results were similar to when considering all planets, with the exception of the planetary radii. The corrected model was still found to be significantly more representative of the data for semi-major axes and orbital period, with Bayes factors of $2.4316 \cdot 10^{5}$ and $3.6704 \cdot 10^{5}$ respectively.

For planetary radii, the Bayes factor was calculated to be $1.4774 \cdot 10^{0}$, meaning there is no strong preference between models rather than the corrected model being highly unfavored as was the case when considering outer planets in our Solar System as well. This improvement of the corrected model is consistent previous analysis, as the outer planets whose detection bias was likely misrepresented are not present in this set of data. The results for stellar radii are trivially unaffected, as they are independent on the number of planets considered.

When plotting the filtered G-type stellar radii 2D histograms against the planets in our solar system we did notice that only Mercury and Venus could be plotted on regions with any sort of data, indicating that even with the corrections, there is more work to be done with detecting exoplanets that are further away from their host star and have a bigger orbital period. It is worth noting that it is possible that since exoplanets have not been detected for an extremely long time, those further away from their host star would be much harder to detect since they could be blocked by their host star for a longer period of time, and the exoplanets researchers simply don't have enough of their data to work with. Also, applying a correction to parameter range that we simply don't have access to as of now, will not be able to properly correct for any bias, since we don't have original data points to work with. It is also possible that the data filtering done at the beginning may have altered some of the data, particularly for the orbital periods and the semi major axes, however the filtering was done to cut off extreme values (there wasn't a slow fall-off, there was simply no data for long ranges before some small peaks in extremely high values).

V. CONCLUSION

In conclusion, we have observed that our corrections were able to mitigate some of the biases introduced by the transit and the radial velocity detection methods, especially for the orbital periods and the semi major axes. Our analysis indicates that the corrected model generally provides a better representation of reality compared to the uncorrected model, except for the parameter of planetary radii where the corrected model appears to be less favored. The discrepancy in the correction for planetary radii suggests that our assumption that in our correction has a flaw, and that parameters like orbital period and radius are not independent of each other. However, in order to properly correct for those biases we would likely need to detect exoplanets for a longer period of time, and get more data for exoplanets that have more similar parameters to the outer solar system planets. We have a lot of exoplanets data, however a lot of them are closer to their host star than Mercury is to the Sun, so our solar system planets are not a representative sample, at least for the data we have now.

In summary, while our corrections have provided valuable insights into the true nature of exoplanets population, there is still much work to be done in understanding the full range of their populations across our galaxy.

Appendix A: Fitted Distributions

[1] N. E. E. Program, Exoplanet exploration program.

[2] R.-K.-U. Heidelberg, Institut für theoretische astrophysik, Retrieved from `https://www.ita.uni-heidelberg.de/~dullemond/lectures/studtage_compastro_2018/Chapter_2.pdf` (2023).

[3] PaulAnthonyWilson.com, The exoplanet transit method (2017).

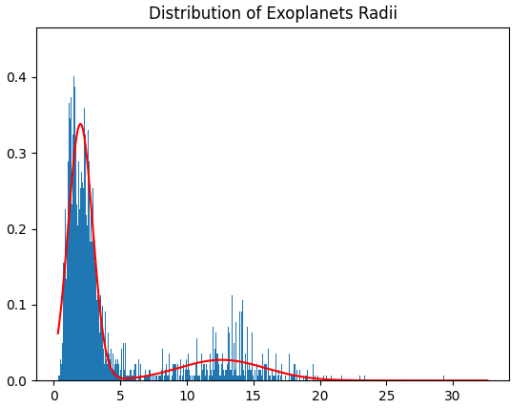
[4] N. E. Archive, Nasa exoplanet archive, (2024).

[5] D. M. Kipping and E. Sandford, Observational biases for transiting planets, Monthly Notices of the Royal Astronomical Society **000**, 1 (2016).

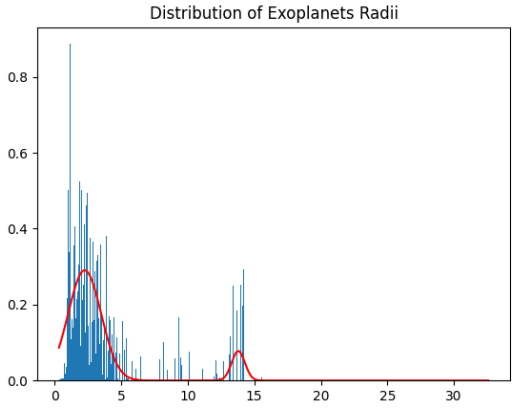
[6] D. R. Williams, Planetary fact sheet, (2024).

[7] S. Conversion, The semimajor axis of planets in our solar system (2023).

[8] C. R. Nave, Kepler's laws, HyperPhysics Mechanics (n.d.).

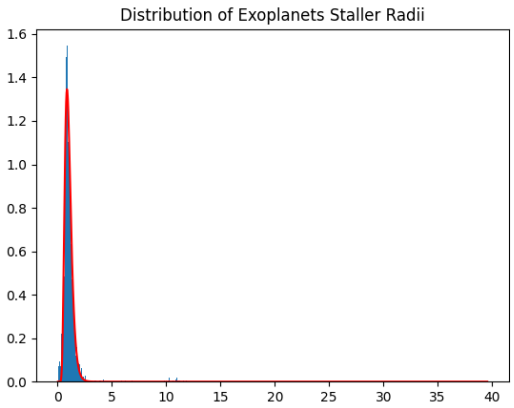


(a) Uncorrected radii distribution fitted with a bimodal distribution

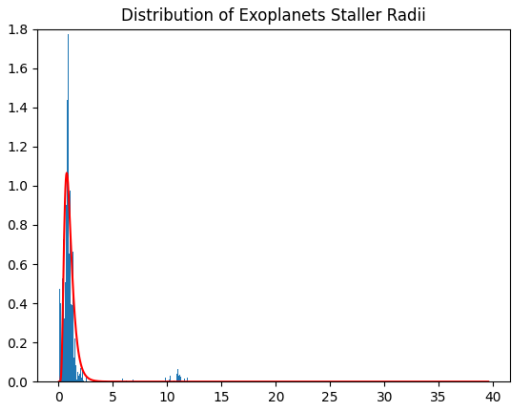


(b) Corrected radii distribution fitted with a bimodal distribution

FIG. 4: Corrected and uncorrected fitting for the radii distributions

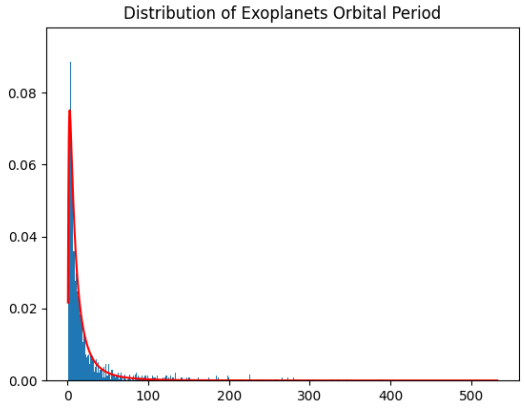


(a) Uncorrected stellar radii distribution fitted with a lognormal distribution

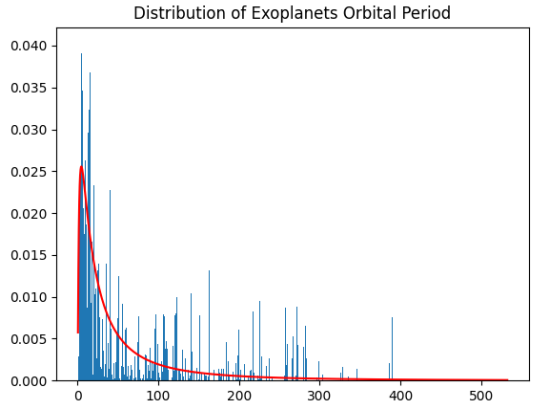


(b) Corrected stellar radii distribution fitted with a lognormal distribution

FIG. 5: Corrected and uncorrected fitting for the stellar radii distributions

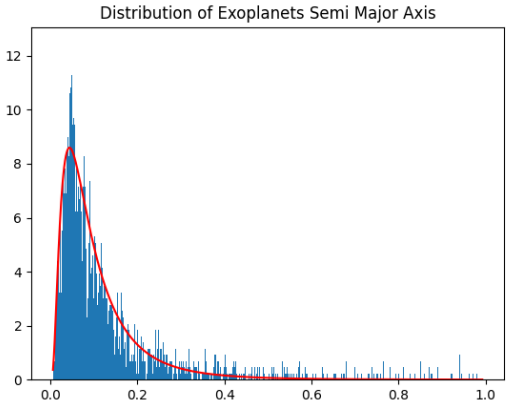


(a) Uncorrected orbital period distribution fitted with a lognormal distribution

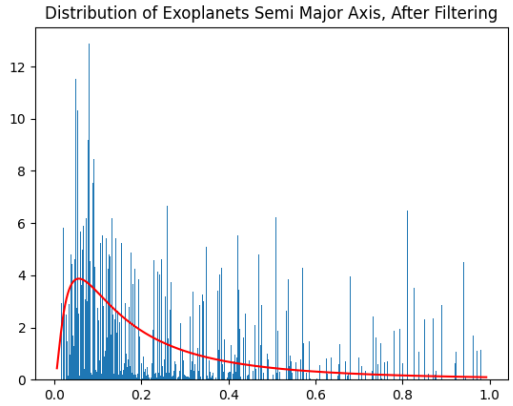


(b) Corrected orbital period distribution fitted with a lognormal distribution

FIG. 6: Corrected and uncorrected fitting for the orbital period distributions



(a) Uncorrected semi major axis distribution fitted with a lognormal distribution



(b) Corrected semi major axis distribution fitted with a lognormal distribution

FIG. 7: Corrected and uncorrected fitting for the semi major axis distributions