



# ASTROCHALLENGE 2020 SENIOR DATA ANALYSIS ROUND

Saturday 5<sup>th</sup> September 2020

**PLEASE READ THESE INSTRUCTIONS CAREFULLY.**

In this part of **ASTROCHALLENGE 2020**, you will work with a moderately large (approx. 4000 points) data set. You will process this data set, analyse it, observe trends, and draw conclusions. **There are no right or wrong answers**; you will be marked solely on the quality of your analysis, even if your statistical methods are incorrect.

We **strongly** recommend you use industry-standard tools like Microsoft Excel™, RStudio or various Python libraries to process the data.

© National University of Singapore Astronomical Society  
© Nanyang Technological University Astronomical Society

# 1 An Introduction to Data Analysis

## Astronomy and Data

Astronomy has always been a data-driven discipline. From the earliest civilisations to the very bleeding edge of 21<sup>st</sup>-century humanity, data has been the core of the matter. Humans have **obtained** and **processed** data; we have **analysed** and **made conclusions** from said data to learn more about the world and the universe we live in.

In other words, astronomy is driven by new observations from data obtained by observing the universe. Scientists employ a very wide variety of techniques and tools to make sense of the data they have obtained. For instance, the CHANDRA X-Ray Observatory observes in X-ray wavelengths, but the data is clearly *not* recorded as X-ray photons! Instead, a computer on board the satellite *encodes* the data from the sensor as bits on some storage, which is then sent to ground-based stations.

This data can then be output into an image, or represented on a graph, or something else. Which representations of data are used, is determined by scientists with respect to the context: whether it be a paper in a scientific journal, or a news article in the morning newspaper, in a popular science magazine, or even as a video on some online platform.

At each step of data analysis, however, lies the possibility for bias to creep in: from the very beginning of data acquisition, to the presentation of data.

This section of **ASTROCHALLENGE 2020** will lead you through the four steps mentioned above, using a live, real data source that can be freely accessed. There are four sections in the paper below. However, this does *not* mean each section corresponds to each step above.

## About the Data Analysis Question

A question *completely* dedicated to data analysis in this form is a complete novelty to **ASTROCHALLENGE 2020**. Hence, you might find some parts to be rather guided—this is intentional. However, others will require you to think out of the box, and be resourceful with the data source.

This question is meant to replace one of the five questions in the Team Round, so the maximum number of marks in this entire round will be **20**. Exact mark allocations will be given within the *right* margin next to each part or sub-part, in boldface and in square brackets, like such: **[2]**.

## Deliverables

This section outlines the deliverables to be submitted for the data analysis question. **Everything** in the following list is to be submitted—any omissions will lead to your team losing marks.

- **Two** .csv files *after* you have completed question **1**;
- A *type-written* report in .pdf format, which should include:
  - Your team’s attempts for all questions;
  - A bibliography, if any external sources were used and cited;
  - An appendix, with all your graphs and images.

## The Data Source

The data source used in this question is *live*—in other words, teams who download their data later may have more (or fewer) data points in their downloaded raw files. This will **not** be a cause for penalty; rather, it merely serves to illuminate the reality of data acquisition in real-life mission: *even when acquiring data from the same source*, the *actual* data acquired may differ between different missions.

## 2 Exoplanets, Exoplanets Everywhere, Nor Any Earth to be Seen

Ever since the detection and subsequent confirmation of the first ever extra-solar planets (exoplanets), PSR B1257+12 B and PSR B1257+12 C, the rate of exoplanet discovery has skyrocketed. The *Kepler* space probe gave new depths to the field, and we now have discovered exoplanets as close as Proxima Centauri. Even this tiny red dwarf, barely larger than Jupiter, hosts one—possibly *two*—planets. Today, we have over 4000 confirmed exoplanets catalogued, and counting.

Complete data on all these exoplanets may be accessed at the [NASA Exoplanet Archive](#), which also has useful tools for data-processing online.

In this round of **ASTROCHALLENGE 2020**, you will process and analyse *all* these data points, and attempt to glean some information from them.


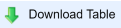
### 2.1 So Many Neighbours...

1. At the end of this question, you ought to have a `.csv`<sup>1</sup> file that is ready to be used directly for the rest of the questions.

- (a) The data source may be accessed at the following link: [http://bit.ly/exopl\\_comp\\_data](http://bit.ly/exopl_comp_data). Before downloading any data, do explore the website, the data source itself, as well as the various table column headers. Explanations of said column headers may be found at the following link: [https://bit.ly/exopl\\_c\\_tbl\\_cols](https://bit.ly/exopl_c_tbl_cols).

This table also corresponds the database column names (given here in monospaced `text`), and the table labels (given here in **boldface**) in the first link above.


To download your data, perform the following steps:

1. Hover your cursor above the Download Table  button, and you should see a drop-down box with several radio buttons.
2. Select the following radio-button options from the drop-down box:
  - CSV Format
  - Download All Columns
  - Download All Rows
3. Click the Download Table  button at the bottom of the drop-down box, to download your `.csv` file.
4. Save this file as `<Your school abbreviation>_T<your team number>_1a.csv`. In other words, if you are in Team 2, Astro Challenge Secondary School, your filename would be `ACSS_T2_1a.csv`.<sup>2</sup> [1]

- (b) Once you have your `.csv` file, you may open it with spreadsheet software<sup>3</sup>. Notice that the data file is prefixed with a large preamble that details, among others, when the file was created, which columns were downloaded, and such. You *may* delete this, but leave the first **two** rows in, which detail when the file was downloaded.

You might realise that the file you downloaded in (a) has more than **360** columns, many of which are *mostly* empty, repeated, or completely unnecessary. You will now trim down the data, and select the most relevant columns.

You may do this in any of the **two** following ways; however, it is recommended that you choose (b)(i). [1]

- (i) In (a), you were tasked to download the data table with **all** columns selected. These columns can be filtered at the website by clicking on the Select Columns button , and checking off boxes on the resulting pop-up.

<sup>1</sup> `.csv` stands for Comma-Separated Values. Such files can actually be opened and manipulated using ordinary text editors like Notepad, TextEdit, Visual Studio Code, Sublime Text, etc. The data are arranged in columns, and separated by commas, which gives rise to the file name.

<sup>2</sup> Please adhere to file name instructions strictly.

<sup>3</sup> such as Microsoft Excel, LibreOffice Calc, Apple Pages, Google Sheets.

Apart from the columns under the `Default Columns` group, all the other checkboxes are initially left unchecked. Refer to **Appendix A** for a list of data columns that you will *require* for the questions below; anything else that you download is optional and no extra credit is given.

Notice that these data columns are arranged rather systematically under collapsible sections, with appropriate headers. Do read these carefully, as well as the rest of the questions below to help you decide which columns you will truly need.

Re-download the new file by following the instructions in (a), except to change the following radio button option:

- Download Currently Checked Columns

Save this new file as `<Your school abbreviation>_T<your team number>_1b.csv`.

- (ii) If you would rather manually edit the `.csv` file, you may do so, too. Refer to **Appendix A** for details on which columns are to be retained, as well as the filename.

## 2.2 I Wonder Who They Are?

2. At this point you will have a fully-workable data file that you can use to plot graphs, calculate regressions, and so on.

The next few questions will take you through basic data processing, and perhaps give you some insight about exoplanetary systems, and some key characteristics about the planets discovered so far.

- (a) (i) Categorise the exoplanets in your data set into the following categories, and present an appropriate diagram. [1]
- |   |   |
|---|---|
| <ul style="list-style-type: none"> <li>• Direct imaging</li> <li>• Eclipse Timing Variations</li> <li>• Gravitational Microlensing</li> </ul> | <ul style="list-style-type: none"> <li>• Radial Velocity</li> <li>• Transit</li> <li>• Transit Timing Variations</li> </ul> |
|---|---|
- (ii) Give reasons for the differences in the number of planets in each category. [2]
- (b) Notice that there are *several* columns for the masses of exoplanets in the dataset: for instance, `p1_bmasse` and `p1_bmassj`. Notice that these masses are labelled **Planet Mass or Mass**  $\times \sin(i)$ . Now, let the symbol  $M_{\text{pl}}$  represent the *true* mass of a given exoplanet.
- (i) Explain what the  $\sin(i)$  refers to, and why despite a multiplication of this factor, both  $M_{\text{pl}}$  and  $M_{\text{pl}} \sin(i)$  may coexist in the same data column. [1]
- (ii) Plot the **planet masses** (choose an appropriate data column) against the **orbital semi-major axis lengths** of the planets (specifically, the data column `p1_orbsmax`).  
**NOTE:** Ensure that your axes are **logarithmic**, and **not linear**. [2]
- (iii) In your plot from (b)(ii), notice that there are three clusters of data points.
- $\alpha$ ) Identify the types of exoplanets that these clusters represent. [1]
- $\beta$ ) If the eight planets in the solar system were plotted on your graph, there would be one cluster which would contain *none* of the planets. Identify and explain which cluster this is. [1]

### 2.3 And What About Their Parents?

3. Notice that the data contains significant information about the exoplanets' parent stars, which can also be processed and analysed.

It is known that the luminosity,  $L$  of a star is related to its mass  $M$ , by the following relationship:

$$\frac{L}{L_{\odot}} = \left( \frac{M}{M_{\odot}} \right)^a \quad (1)$$

where  $L_{\odot}$  and  $M_{\odot}$  refer to the solar luminosity and mass, respectively, and  $a$  is some exponent.

Using a scatter-plot graph, determine the value of the exponent  $a$  and its corresponding uncertainty for star masses  $M$  in the range  $0.43 M_{\odot} \leq M \leq 2 M_{\odot}$ . [3]

### 2.4 Are We Alone?

4. A key reason for the search for exoplanets is to find another exoplanet that is *very* similar to Earth, and hence, potentially habitable by life as we know it. Some measures of *habitability* and *similarity* to Earth have been proposed by scientists. These scales typically output a *single* number,  $x$ , where  $0 \leq x \leq 1$ . On such a scale, 0 is a planet that is as far removed from similarity or habitability to Earth as possible, and 1 is Earth itself.

Two such indices are the Earth Similarity Index (ESI), and the Planetary Habitability Index (PHI)<sup>4</sup>.

- (a) The ESI for a planet (extrasolar or otherwise), and for a given physical property  $i$  of said planet out of  $n$  physical properties, is defined as follows:

$$\text{ESI}_i = \left( 1 - \left| \frac{x_i - x_{i0}}{x_i + x_{i0}} \right| \right)^{\frac{w_i}{n}}$$

$$\text{ESI} = \prod_{i=1}^n \text{ESI}_i \quad (2)$$

where  $x_i$  and  $x_{i0}$  are the above-mentioned physical properties of the planet and Earth respectively, such as radius, density, etc;  $w_i$  is a weighted exponent, and  $n$  is the total number of properties.

One such combination of *four* physical factors gives rise to an ESI defined by Schulze-Makuch et al.:

$$\text{ESI} = \left( \text{ESI}_r \cdot \text{ESI}_{\rho} \cdot \text{ESI}_{v_e} \cdot \text{ESI}_{T_s} \right)^{\frac{1}{4}} \quad (3)$$

where  $r$  is the radius,  $\rho$  is the bulk density,  $v_e$  is the escape velocity, and  $T_s$  is the mean temperature.

The ESI value is *fairly* comprehensive, as it accounts for *interior* similarity ( $r$  and  $\rho$ ), as well as *surface* similarity ( $v_e$  and  $T_s$ ). However, a planet that achieves a near-Earth ESI score (e.g.  $\text{ESI} \geq 0.90$ ) might still be *completely* uninhabitable. This implies that the ESI as defined in **equation (3)** excludes certain critical parameters that allow life to develop and flourish.

List some planetary parameters that might be missing in the ESI, and explain their significance. [2]

- (b) A second measure, meant to determine the *habitability* of a certain planet and not merely its physical similarity to Earth, has been proposed. This measure is called the Planetary Habitability Index (PHI). The PHI has a *much* more complex definition than the ESI, the former of which may be found in full, in the reference cited. A simplified definition is shown below, in **equation (4)**.

$$\text{PHI} = (S \cdot E \cdot C \cdot L)^{\frac{1}{4}} \quad (4)$$

<sup>4</sup>Dirk Schulze-Makuch et al. 'A Two-Tiered Approach to Assessing the Habitability of Exoplanets'. In: *Astrobiology* 11.10 (21st Oct. 2011), pp. 1041–1052. ISSN: 1531-1074. DOI: [10.1089/ast.2010.0592](https://doi.org/10.1089/ast.2010.0592). URL: <https://www.liebertpub.com/doi/abs/10.1089/ast.2010.0592> (visited on 23/08/2020).

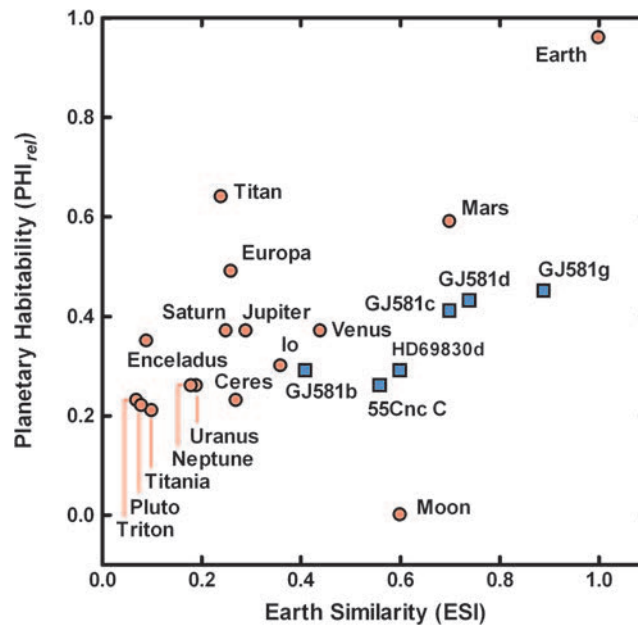
*S* refers to *substrate*: presence of a solid surface (rocky or icy), presence and thickness of an atmosphere, and magnetosphere strength are counted.

*E* refers to *energy*: distance to the parent star, the mean surface temperature, and tidal flexing/locking are considered for this measure.

*C* refers to *chemistry*: presence of complex polymeric, or simple organic compounds are factors.

Finally, *L* refers to *liquids*: presence of liquids on the surface of the planet contribute to this.

The PHI is *normalised* by dividing all values by the largest value in the set of PHI values of planets: this yields a *relative* value  $\text{PHI}_{\text{rel}}$  between 0 and 1, similar to the ESI.



**Figure 1:** Values for the global ESI and the relative PHI ( $\text{PHI}_{\text{rel}}$ ) of solar system bodies (red circles), and selected exoplanets (blue squares).<sup>5</sup>

Refer to **Figure 1** above. Explain why the plotted points do not *all* fall on the line  $y = x$ . In other words, explain why a given planet's PHI may not equal its ESI value, or vice versa. [2]

## 2.5 Why Are We Here?

5. Humanity's search for exoplanets (habitable or otherwise) may be reduced to a straightforward problem of achieving a greater sample size from an immensely large population. In this case, the *population* is the set of *all* exoplanets currently in existence in the observable universe; certain estimates of this number exceed several billion trillion (i.e.  $10^{21}$ ).

The current sample size is simply the number of exoplanets we have discovered so far; as you may have noted, this number is *approximately* 4000, and steadily increasing as past data from active and defunct missions is analysed, and new data is acquired from existing missions. The Transiting Exoplanet Survey Satellite (TESS) and the James Webb Space Telescope (JWST) are set to revolutionise exoplanet surveying, just like *Kepler* did.

However, much of the dataset provided is probably rife with errors and biases. Filtering out the bad data from the good, uses up man-hours and computational capacity, which could have been otherwise used to determine if an exoplanetary candidate is valid or not.

Give some sources of errors and biases in the data that you have downloaded in question 1(b). Explain why you think these sources are significant in their error/bias contribution. [3]

## A Data Columns

This appendix details data column codes that you will **mandatorily** need in this round; ensure you have selected these when attempting question **1(b)** and onwards.

Database Column Name	Table Label	Description
pl_hostname	Host Star Name	Stellar name most commonly used in the literature.
pl_letter	Planet Letter	Letter assigned to the planetary component of a planetary system.
pl_discmethod	Discovery Method	Method by which the planet was first identified.
pl_pnum	Number of Planets in System	Number of planets in the planetary system.
pl_orbper	Orbital Period (days)	Time the planet takes to make a complete orbit around the host star or system.
pl_orbsmax	Orbit Semi-Major Axis (AU)	The longest radius of an elliptic orbit.
pl_orbeccen	Eccentricity	Amount by which the orbit of the planet deviates from a perfect circle.
pl_orbincl	Inclination (°)	Angular distance of the orbital plane from the line of sight.
pl_massj	Planet Mass ( $M_J$ )	Amount of matter contained in the planet, measured in units of masses of Jupiter.
pl_msinij	Planet $M \cdot \sin(i)$ ( $M_J$ )	Minimum mass of a planet as measured by radial velocity, measured in units of masses of Jupiter.
pl_radj	Planet Radius (Jupiter radii)	Length of a line segment from the center of the planet to its surface, measured in units of radius of Jupiter.
pl_dens	Planet Density ( $\text{g cm}^{-3}$ )	Amount of mass per unit of volume of the planet.
st_dist	Distance (pc)	Distance to the planetary system in units of parsecs.
st_teff	Effective Temperature (K)	Temperature of the star as modeled by a black body emitting the same total amount of electromagnetic radiation.
st_mass	Stellar Mass ( $M_\odot$ )	Amount of matter contained in the star, measured in units of masses of the Sun.
st_rad	Stellar Radius (solar radii)	Length of a line segment from the center of the star to its surface, measured in units of radius of the Sun.
pl_name	Planet Name	Planet name most commonly used in the literature.
pl_masse	Planet Mass ( $M_\oplus$ )	Amount of matter contained in the planet, measured in units of masses of the Earth.
pl_msinie	Planet $M \cdot \sin(i)$ ( $M_\oplus$ )	Minimum mass of a planet as measured by radial velocity, measured in units of masses of Earth.
pl_rade	Planet Radius (Earth radii)	Length of a line segment from the center of the planet to its surface, measured in units of radius of the Earth.
st_lum	Luminosity [ $\log(\text{solar})$ ]	Amount of energy emitted by a star per unit time, measured in units of solar luminosities.