Data Science in Astrophysics 2024:
Worksheet 1: Statistics & Spatial Statistics

You should submit your solutions to this task as a worked out report (PDF), along with the code you used to solve the tasks (.py or .ipynb). The report should contain

- a brief introduction,
- the results of your analysis (structured according to the tasks) along with all plots,
- a brief description of your implementation, and
- a brief discussion of the results.

The report should be written in a clear and concise manner, but it should contain enough information to allow the reader to understand what you did and why. It is not enough to just submit plots. The code should be well-commented and structured in a way that makes it easy to understand what it does.

Since there are many scatter plots to be made, export them to a rasterized image file (e.g. PNG) rather than a PDF to keep the file size manageable.

Project: The spatial distribution of SDSS galaxies

Getting the data

Download the file `sdss_cutout.csv` from the Moodle page. It is a simple CSV (comma separated values) file with the columns: RA, DEC, redshift z, u magnitude, g magnitude, r magnitude, i magnitude, z magnitude.

You can load it into Python using either numpy:

```
1  import numpy as np
2  d = np.loadtxt('sdss_cutout.csv', skiprows=1, delimiter=',')
```

and then access the columns via array index, e.g. `RA = d[...,0]`. Alternatively, you can use the Pandas framework, to load the data into a data frame

```
3  import pandas as pd
4  d = pd.read_csv('sdss_cutout.csv')
```

in which case the data is accessible via dictionary access, e.g. `mag_i = d['phot_i']`.

When you have loaded the data, you should have around 320'000 galaxies from the SDSS galaxy survey.

Dataset background knowledge (for the non-astronomers)

The Sloan Digital Sky Survey (SDSS). The SDSS is a large survey of galaxies, quasars and stars. It has imaged more than a quarter of the sky and has measured the redshifts of more than a million galaxies. The SDSS has been used to study the large-scale structure of the universe, the distribution of galaxies, the properties of the intergalactic medium, and the properties of the Milky Way.

Sky coordinates The position of a galaxy on the sky is given in terms of its right ascension (RA) and declination (DEC), measured in degrees.

Cosmic expansion and redshift. Due to the cosmic expansion, the light from distant galaxies is redshifted. According to the Hubble law, there is a one-to-one relation between the distance and the redshift z of a galaxy, where more distant galaxies have a higher redshift z. However, the redshift of a galaxy is also affected by its peculiar velocity, i.e. its motion relative to the cosmic expansion. Therefore, the observed redshift of a galaxy is a combination of the cosmological redshift and its peculiar velocity. The peculiar velocity is typically much smaller than the cosmological redshift, but it can be significant for nearby galaxies. For the galaxies in the SDSS, redshifts have been measured by observing the shift of spectral lines. We will be neglecting the role of peculiar velocities in this task, the redshift is therefore simply an indicator of cosmological distance.

Magnitudes. The luminosity of a galaxy is measured in various filter-bands of the electromagnetic spectrum. The measured magnitudes are apparent magnitudes and a galaxy of fixed intrinsic luminosity will appear fainter with increasing distance. The SDSS is magnitude limited, i.e. it only contains galaxies that are bright enough to be observed. Nearby galaxies are therefore observed to a lower intrinsic luminosity than distant galaxies, while only the brightest galaxies are included at high redshift. The intrinsic luminosity of a galaxy is related to its stellar mass. Magnitude is related to luminosity by the relation $m = -2.5 \log_{10} L + \text{const}$, where m is the magnitude and L is the luminosity (i.e. fainter = larger number). The SDSS measures the apparent magnitude of galaxies in five bands: u, g, r, i, and z.

Colors. Galaxy colors are the difference in magnitude between two bands. The color of a galaxy is related to its stellar population. For example, a galaxy with a young stellar population will have a different color than a galaxy with an old stellar population. The color of a galaxy is also related to its star formation rate. Typically red galaxies have an old stellar population and a low star formation rate, while blue galaxies have a young stellar population and a high star formation rate. We also know that massive galaxies are typically red and low-mass galaxies are typically blue. Note however that due to

the cosmological redshift, the observed colors of galaxies are also affected by the redshift and do not agree with the intrinsic colors of the galaxies.

Exploring the data

Task 1. Plot the eCDF of the redshift distribution of the catalog. Note that the catalog is (apparent) magnitude limited (magnitude $r < 17.7$, i.e. fainter objects with $r > 17.7$ are not detected). Make a scatter plot of r-band magnitude vs. redshift (make sure you use small points and give them a small alpha value, e.g. `plt.scatter(xdata, ydata, s=0.5, alpha=0.1`). Discuss the competing effects of increasing volume vs. the effect of the magnitude cut with increasing z (i.e. what does it mean if we want a 'complete' sample). Discuss why it seems reasonable to focus only on the redshift range $0.08 < z < 0.12$. Apply this redshift cut to the data for all the analysis that follows.

Task 2. Make a color-magnitude scatter plot of the r-band magnitude vs. the $u - r$ color for the galaxies. Show that the red and blue galaxies are well separated in this color-magnitude diagram. Split the galaxy population into red vs. blue galaxies by computing for each galaxy the photometric color $u - r$ and selecting a 'red' sample with $u - r > 2.3$ and a blue sample with $u - r \leq 2.3$.

Task 3. Compute the mean and standard deviation of the r-band magnitude for the red and blue galaxies. Discuss the results (i.e. what do you learn about the luminosity of red vs. blue galaxies? are they substantially different?).

Task 4. Create scatter plots of RA vs. DEC (the 'angular map') and RA vs. redshift z (the 'redshift-space map') for both galaxy samples (make sure you use small points and give them a small alpha value, e.g. `plt.scatter(xdata, ydata, s=0.5, alpha=0.1)`). Discuss the qualitative differences you observe.

Computing the two-point correlation function

Task 5. Write PYTHON code to estimate the angular two-point (auto-)correlation function $\hat{\xi}_{\rm red,red}(\theta)$ of the red galaxies and $\hat{\xi}_{\rm blue,blue}(\theta)$ of the blue galaxies using the Landy-Szalay estimator, where θ is the angle between a galaxy with (RA,DEC) (α_1, δ_1) and another at (α_2, δ_2). To do this, you need to write an efficient PYTHON code that

1. Make a realisation (by drawing and rescaling uniformly distributed random numbers from `np.random.rand(M)`) of the M random points with the same extent in RA-DEC space $130 < \alpha/\text{degrees} < 230$ and $5 < \delta/\text{degrees} < 65$. Normally these should not be drawn uniformly in α and δ due to the sphere surface metric, but let us neglect this here. [For precision measurements, one uses $M \sim 10N$ but this is too much for this test, use $M \approx N$.]

2. Create 10 logarithmically space angular separation bins
`omega = np.geomspace(0.003,0.3,11)` (these are the limits of the bins in radians).
3. Count the number of pairs $DD(\omega)$, $RR(\omega)$ and $DR(\omega)$ that fall in each of the bins and use the Landy-Szalay estimator to obtain an estimate of the correlation function.
4. Present the plots (in log-log) of the correlation function of blue and red galaxies. Can you connect what you see here with the scatter plots of the positions of the galaxies you made before?

Task 6. Add a confidence interval to the correlation function. To do this, you need to estimate the error of the mean of the correlation function. This can be done in multiple ways, e.g. by delete-d jackknife (compute the correlation function multiple times for subsamples, then compute the variance over the subsamples). [use few sub-samples to keep the computation time manageable].

Task 7. Discuss briefly the results. Find information on the internet about the expected behavior of the two-point correlation function (aka the clustering) of red and blue galaxies. Do your results qualitatively agree with this? What do you learn about the spatial distribution of red and blue galaxies?

Hint 1: You can convert between RA,DEC and spherical coordinates using

$$\phi = \pi \frac{\mathrm{RA}}{180} \qquad \theta = \frac{\pi}{2} - \pi \frac{\mathrm{DEC}}{180}$$

Once you have that, the angle between two points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ on the unit sphere is

$$\omega = \arccos \boldsymbol{x}_1 \cdot \boldsymbol{x}_2 = \arccos \left(\cos\theta_1 \cos\theta_2 + \cos(\phi_1 - \phi_2) \sin\theta_1 \sin\theta_2\right)$$

Hint 2: you need to do this avoiding explicitly looping over all pairs in PYTHON. *At the same time, you cannot compute a distance matrix (it is too big). Either use* NUMBA *when looping, or divide your sample into batches and compute the matrix of pairwise distances for the batch, while looping over all combinations of batches.*

Hint 3: if your laptop is not fast enough and/or for code development, randomly downsample the data sets. The improvement in speed is quadratic (i.e. half the data, four times as fast).