Homework 01

ASTROSTATISTICS

1. Take a normal distribution with mean equals to 100 and variance equals to 10, i.e., $\mathcal{N}(100, \sqrt{10})$. Produce N samples with 8 data points each assuming such a distribution, and prove the central limit theorem using those samples.

2. Consider now the DR12 quasar (QSO) catalogue from the Sloan Digital Sky Survey **(SDSS)**, which comprises 297,301 uniquely identified QSOs

a Download the catalogue from the SDSS website – or any other database. State where the dataset was obtained from.

b Plot a histogram of the redshift distribution of the objects in this catalogue. State what is the bin width assumed in these plots.

c Fit a Gaussian curve to this histogram, and compute the mean, variance and higher order momenta (skewness and kurtosis). Would you consider that this Gaussian PDF is a good fit to the data? Explain. *Tip*: You can use PYTHON fitting routines to perform this procedure.

d Instead of a Gaussian curve, try to fit a Gamma function to this histogram, i.e.,

$$P(z;\ k, \theta) = z^{k-1}\frac{e^{-z/\theta}}{\theta^k \Gamma(k)} \tag{1}$$

where z is the redshift, while k and θ are the free parameters of the Gamma function. What are the best fits for those parameters? Moreover, which one provides a better fit to the data, the Gamma or the Gaussian curve? Explain.

3. Consider now the Type Ia Supernova (SN) compilation named `Union2.1` sent in the mail. Also consider that the file columns represent SN name, redshift, distance modulus and the distance modulus uncertainty, respectively. You can neglect the last column in the meantime.

a Given that the distance modulus follows $\mu \equiv m - M = 5\log d - 5$, where d denotes the distance to the source in Pc (parsec) unit, show that $\mu = 5\log d_L + 25$, for a luminosity distance d_L given in Mpc (Megaparsec).

b In the standard cosmological model, namely the flat ΛCDM paradigm, the luminosity distance reads

$$d_L = cH_0^{-1}(1+z)\int_0^z \frac{dz'}{\sqrt{\Omega_m(1+z')^3 + (1-\Omega_m)}}\,, \tag{2}$$

where Ω_m denotes the total matter density parameter of the Universe. By fixing the Hubble Constant $H_0 = 70$ km s^{-1} Mpc^{-1}, compute the χ^2 for $\Omega_m = 0.20$, $\Omega_m = 0.30$, and $\Omega_m = 0.40$, which reads

$$\chi^2 = \sum_i \frac{(\mu^{model}(z_i;\ \Omega_m, H_0) - \mu_i^{data})^2}{\sigma_i^2}\,. \tag{3}$$

In this case, μ^{model} and μ^{data} consist on the distance modulus assuming the standard cosmological model and the observational data measurement, respectively, σ denotes the uncertainty of the μ^{data} measure, z is the observational redshift, and the i subscript corresponds to the i-th data point of the observational sample. What Ω_m value provided the best fit among them all – in other words, the smallest χ^2?

c In statistical inference, a quick way to verify whether the fit of a parameter is good or not is the so-called reduced χ^2, hereafter χ^2_ν, which is expressed as the χ^2 divided by the number of degrees of freedom – so the number of data points minus the number of constraints, i.e., the numbers of parameters in a model. Given the values obtained for the latter quantity, what is the Ω_m value that gives the best fit to the data? Explain. *Tip:* Note that $\chi^2_\nu \sim 1$ is a good indicator for a reasonable fit. Values above (below) this threshold may indicate underfitting (overfitting).

d Compute the best fit value overall for Ω_m, rather than fixing it at specific values, using the χ^2 as in Eq. (3). Make a plot of the χ^2 in function of $\Omega_{\rm m}$, as well as the likelihood, according to

$$\mathcal{L} \propto \exp\left(-\chi^2/2\right). \tag{4}$$

What happens if you plot $\mathcal{L}/\mathcal{L}_{max}$? Explain.

e Fit a Gaussian curve to the likelihood obtained in item [d], and compute the mean and standard variation from this fit. Does the mean coincide with the best-fitted value obtained in it? Also explain.

4. Consider the data provided in the DATA_HMWRK01_EX4.DAT file. Let us assume that these value represent the number counts of Active Galaxy Nuclei (AGNs) across 20 disjoint sky regions of the same angular size. Perform a resampling procedure using the BOOTSTRAP method with $N = 100$, $N = 1000$ and $N = 10000$. What is the mean and standard deviation obtained from them? Plot the results and explain.

5. Do an arXiv search of "[research field] + "distribution [name of the distribution]", among those we have covered so far – e.g. "gravitational wave" + "binomial distribution", "radio-loud AGNs" + "gaussian distribution" etc. You can choose make as many different searches as you wish. Show the link of the papers you found with those searches, and make some brief comments about the results reported in those papers as well.