Statistics for Astronomers
Midterm Exam (Due before 5:00 PM on Tuesday, 2021.02.09)

Prof. Sundar Srinivasan

February 8, 2021

Notes: (1) You are welcome to use `Python` functions to evaluate probabilities for various distributions, and `Mathematica/Wolfram Alpha` to compute integrals if necessary. **Just mention your source in each case!** (2) Email me your `Python` scripts and any/all resulting output plots/images.

1. (**6 points**) The file SN_data.vot (VOTable) contains the morphology type and stellar mass of the host galaxies of 68 Type Ia supernovae. Perform a t-test to determine whether galaxies with types S0 or earlier have larger stellar masses than those with types S0/a or later. Repeat the test for (S0/a or earlier) vs. (Sa or later), and for (Sab or earlier) vs. (Sb or later).

   Note: assume in each case that the two subgroups are drawn from populations with differing variances.

2. Rishi wants to compute a linear fit to his data (VOTable) consisting of observations (x, y) with measurement uncertainties sy on the y variable. He performs a χ^2 minimisation procedure that results in best-fit estimates of 27 and -10.4 for the intercept and slope respectively.

   (a) (**3 points**) Rishi considers χ^2 values within the 68% central confidence interval around the expected χ^2 for his dataset as "acceptable". For the best-fit parameter values given above, does he have an acceptable fit?

   (b) (**5 points**) Sarah suspects that the data might violate some of the conditions required for the χ^2 distribution to be applicable. Design an appropriate hypothesis test and determine whether her suspicions are valid.

3. In this problem, you will compute a power-law relation for the excess flux at 8 μm for LMC C-rich AGB stars as a function of their luminosity. The data required for this problem can be downloaded here (CSV). Given the luminosity L in solar luminosities and the 8 μm excess flux F_8 in Jy, you have to find parameters (α, β) such that

   $$y = \alpha + \beta\ x, \qquad \text{with } x \equiv \log L, y \equiv \log F_8. \tag{1}$$

   Assume that the uncertainties associated with L and F_8 are independent, uncorrelated, and normally distributed.

   It will help to visualise the data on a log-log plot before you answer the questions that follow.

   (a) (**3 points**) Propagate the uncertainties in (L, F_8) to the uncertainties (s_x, s_y) in (x, y). You can use the linear approximation of the Taylor Series ($i.e$, just use the first derivative).

(b) (**1 point**) Using the (x, y) values and the uncertainty s_y you obtained from the previous parts, fit a straight line using the method described in Section 1 of Hogg et al. (2010). What are the intercept and the slope?

(c) (**1 point**) Based on the resulting covariance matrix for the parameters, what are the uncertainties in the parameters?

(d) (**2 points**) What is the correlation coefficient between the uncertainties in the intercept and the slope?

(e) (**2 points**) Compute the **reduced** χ^2 using the (x, y) values, the uncertainty s_y, and the best-fit intercept and slope (*Hint: use Equation 7 from Hogg et al. and divide the* χ^2 *by the number of degrees of freedom*). Is the value very different from unity? If so, what do you think it is the reason?

(f) (**1 point**) Based on the discussion in Section 4 of Hogg et al., are the parameter uncertainties computed in Question 3c realistic? Why/why not?

(g) (**5 points**) Use $B = 100$ bootstrap resamples to estimate the standard deviations for the intercept and the slope. **Caution: this is a time-consuming step, so make sure this part of the code runs independent of the rest.**

4. We will now improve the fit to the data in Question 3 using the description in Sections 7 and 8 in Hogg et al. In these sections, the paper describes a method to fit a line by incorporating uncertainties along both axes as well as intrinsic scatter. In this method, we transform the intercept and the slope into parameters $\theta \equiv \tan^{-1}(\text{slope})$ (the angle subtended by the line at the X-axis) and $b_\perp \equiv \text{intercept}/\cos\theta$ (the perpendicular distance of this line from the origin). An additional parameter V accounts for intrinsic scatter **orthogonal to the line** (V is the variance of the orthogonal intrinsic scatter).

   Download `orthofit.py`. This code performs maximum likelihood estimation using Equation (35) in Hogg et al. to derive the best-fit values for θ, $b_\perp$, and V. In order to perform the fitting, the code requires the data (x, y) and the uncertainties (s_x, s_y) from Question 3 as input. It also requires an initial guess vector for the parameters $\theta, b_\perp$, and V. The code then outputs the best-fit values for the intercept and slope (transforming back from $\theta, b_\perp$), and V.

   (a) (**1 point**) Use the best-fit intercept and slope computed in Question 3b to derive initial guesses for θ and $b_\perp$.

   (b) (**2 points**) In Question 3e, the reduced χ^2 was computed assuming that the covariance matrix only had contributions from s_y. Suppose instead that the covariance matrix was of the form
   `Sigma = np.diag(s_x**2 + s_y**2 + invar)`,
   where `invar` is a number less than 1. From trial-and-error, find any one value of `invar` for which the reduced χ^2 is close to 1 (remember, the number of parameters has increased by 1 because of `invar`). Use this value of `invar` as the initial guess for V.

   (c) (**4 points**) Execute `orthofit.py`. It outputs the best-fit intercept, slope, and intrinsic variance. Plot the data onto a figure and overlay a line generated from the (intercept, slope) pair computed in this question, and compare it to a line generated from the pair computed in Question 3b.