Statistics for Astronomers
Final Exam (09:00 Thursday, 2019.05.30 – 17:00 Friday, 2019.05.31)

Prof. Sundar Srinivasan

Instructions:
– You are free to use the course notes and solutions, as well as any online material. I only require that you cite your sources for such materials if any.
– Some scripts from your homework submissions will be useful for this exam. You're welcome to paste/import them into code for the exam.
– As usual, submit your scripts/plots electronically and submit any written material in person before the deadline.
– Most importantly, don't panic. Email me if you have any doubts. I'll also be in the office all day Friday.

Questions

1. A company has announced that some bottles of its flagship soft drink, Ískaldur, have a code printed on them that entitles the bearer to an all-expenses paid trip to Iceland. The probability that any bottle is a "winner" is $1/N$. You purchase N bottles of Ískaldur.

   (a) (**2 points**) What is the probability that at least one of the bottles you purchased has a winning coupon code?

   (b) (**1 point**) What is the value of this probability as $N \to \infty$?

2. (a) (**3 points**) If $u \sim U(0, 1)$ and $t = u^{\alpha}$, then what is the pdf of t?

   (b) (**2 points**) Using the above result, how would you use uniform random numbers to draw from the distribution $p_T(t) = \frac{1}{3}t^{-2/3}$, with $0 < t \leq 1$?

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. 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

$$F_8 = \alpha L^{\beta}. \tag{1}$$

   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**) Is Equation 1 linear in the parameters (α, β)? If not, define new variables (x, y) and parameters (a, b) such that you transform Equation 1 into the equation for a line.

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

(c) (**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?

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

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

(f) (**2 points**) Compute the **reduced** χ^2 using the (x, y) values, the uncertainty s_y, and the best-fit intercept and slope. Is the value very different from unity? If so, what do you think it is the reason?

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

(h) (**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. Sections 7 and 8 in Hogg et al. describe one method to fit a line that incorporates uncertainties along both axes as well as intrinsic scatter. In this method, we transform the intercept and the slope into parameters θ (the angle subtended by the line at the X-axis) and $b_\perp$ (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). It also requires an initial guess vector for the parameters $\theta, b_\perp$, and V. The code then outputs the best-fit values for b, m (transforming back from $\theta, b_\perp$), and V.

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

   (b) (**2 points**) In Question 3f, 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 3c.

   (d) (**2 points**) In the above, we started with a likelihood defined for three parameters (the intercept and slope defined through θ and $b_\perp$, and the intrinsic variance orthogonal to the best-fit line) and maximised it. In your own words, explain how you would proceed if, instead of the above procedure, you decided to "go Bayesian".