Statistics for Astronomers
Midterm Exam (Due before 5:00 PM on Tuesday, 2020.11.10)

Prof. Sundar Srinivasan

November 9, 2020

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. Two random variables, X and Y, have a joint pdf given by

$$p_{XY}(x, y) = e^{-y}, \quad 0 < x < y < \infty, \quad \text{zero elsewhere.}$$

   (a) (**2 points**) Determine the marginal pdf of x and the marginal pdf of y.

   (b) (**2 points**) Determine the conditional pdf of x given y and of y given x.

   (c) (**1 point**) Are x and y independent random variables? Why (not)?

2. If you are only allowed to draw random numbers from $U(0, 1)$, describe how you would generate random numbers distributed according to

   (a) (**3 points**) $p_T(t) = \frac{1}{3}\, t^{-2/3}, \quad 0 < t \leq 1$ and

   (b) (**5 points**) $p_X(x) \propto \left(\frac{x^2}{\nu} + 1\right)^{-1}, \quad -\infty < x < \infty$

3. (**3 points**) A data set with 25 observations has sample mean m and sample standard deviation s. If it is drawn from a distribution with population mean μ, what is the distribution of $\frac{m - \mu}{s}$? What are its expectation value and standard deviation?

4. (**4 points**) N data points are drawn from the distribution $p_X(x) \propto e^{-|x-\mu|}, \ -\infty < x < \infty$. Compute the MLE for μ.

5. (**7 points**) The file linked here contains 50 000 samples from the likelihood distribution of a parameter. Plot the distribution and its CDF (the empirical distribution function), then write a script that returns the lower and upper limits of the 20% shortest CI for the true value of the parameter.

6. (**4 points**) Given that the mean discovery rate of Type IIn supernovae is 7 yr^{-1}. Write a code to predict the number of supernovae to be discovered in 2021 – produce a **symmetric CI about the expectation value** with confidence as close to 75% as possible (it won't be exact since the number of events is an integer).

7. The file linked here contains data for the hypothetical spectrum shown in Figure 1. The data is in terms of photon counts as a function of the wavelength bin for 15 bins. The seven central bins (the "line region") contain a spectral line overlaid on the continuum. Write a Python script to answer the following questions. Where applicable, have your script print out the result.

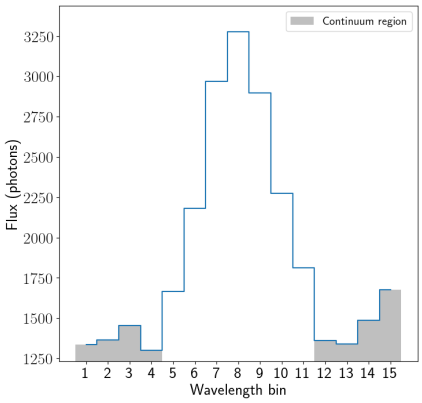

Figure 1: Hypothetical spectrum for Problems 7 and 8. The first four bins on either side are to be used to estimate the continuum.

(a) (**3 points**) Use the bins on either side of the line region to estimate the mean continuum count per bin, and the standard error (square root of the variance) associated with it.

(b) (**3 points**) Use the mean continuum count to estimate the total continuum contribution under the line, and its standard error.

(c) (**4 points**) Subtract the continuum contribution from the total counts in the line region to calculate the line contribution. Also compute the associated standard error by combining the errors in the continuum contribution and that from the total counts.

(d) (**1 points**) In terms of data quality, what is the advantage of reporting the total counts in the line rather than the count at the line centre (in the central bin)?

8. Suppose now that, for the spectrum in Question 7, the count in a bin is strongly correlated with the counts in its nearest-neighbour bins. That is,

$$\mathrm{Cov}[C_i, C_j] = \rho_{ij}\ \sqrt{\mathrm{Var}[C_i]\ \mathrm{Var}[C_j]}, \quad \text{with } \rho_{ij} = \begin{cases} 1, & \text{if } i = j \\ 0.5, & \text{if } |i-j| = 1 \\ 0, & \text{otherwise} \end{cases}$$

(a) (**4 points**) Compute the covariance matrix for the counts in the first four bins (*i.e.*, the bins to the left of the line region).

(b) (**4 points**) With the nearest-neighbour correlated bins, how does your answer to 7a change?