Statistics for Astronomers
Midterm Examination (Monday, 2019.03.25, 09:00 – 12:00)

Prof. Sundar Srinivasan

Instructions:

1. You have three hours to attempt as many questions as possible.
2. For questions that require programming, email me your code and the resulting output and/or plots.
3. For questions that do not require programming, you can leave your numerical answers in the form of fractions and/or in terms of the information provided below.

Useful information:

1. $e^{-1/2} = 0.606530$, and $e^{-25/2} \approx 4 \times 10^{-6}$.
2. If $\Phi(y)$ is the CDF of the standard normal distribution,
$$\Phi(-1) = 0.1586552, \Phi(3) \approx 0.997, \Phi(5) = 0.9999997, \text{ and } \Phi^{-1}\left(\frac{1}{2}\Big[\Phi(-1) + \Phi(5)\Big]\right) = 0.200.$$

Questions

1. **(5 points)** Assume that stars of the same spectral type have the same radial speed $v_{\rm rad}$, but that their directions are oriented randomly. Thus, the projected radial velocities are $v_{\rm rad} \sin\phi$, where ϕ (the angle between the line-of-sight and the radial velocity vector) is drawn from $U[0, \pi)$. Find **(a)** the probability distribution of the projected velocities, **(b)** the population mean and **(c)** the population standard deviation.

2. **(3 points)**
A 100-seater plane has a passenger load limit of 8450 kg. Assuming that the passenger masses are independent and identically distributed according to $\mathcal{N}(\mu, \sigma^2)$, with $\mu = 80$ kg and $\sigma = 15$ kg, what is the probability that the load limit is exceeded?

3. **(8 points)**
A teacher discovers that the final grades for a class follow a distribution of the form
$$p_X(x) = \begin{cases} C \exp\left[-\dfrac{1}{2}\left(\dfrac{x-\mu}{\sigma}\right)^2\right] & \mu - \sigma \leq x \leq \mu + 5\sigma \\ 0 & \text{otherwise,} \end{cases}$$
with constants C, μ, and σ.

   (a) In terms of μ and σ, what are the mean and median scores?

(b) What is the physical interpretation of the parameter μ?

(c) Given the scores x_i for the N students in the class, what is the maximum likelihood estimate for μ? Assume that one student's score is independent of the scores of the other students.

(d) What is the expectation value of the maximum likelihood estimator computed in (3c)? Is it an unbiased estimator for μ?

4. **(3 points)**
It is estimated (*e.g.*, Hair et al. 2013) that roughly 10^{-5} of transient radio signals detected might originate from extraterrestrial civilisations. An algorithm that classifies these transients as either "N" (for natural) or "ET" (for extraterrestrial) has Type-I and -II error rates of 0.1% each. In such a case, what fraction of the signals classified as "ET" are actually artificial?

5. **(1 point)**
Given a sample of 10 data values drawn from a population, we are required to place confidence limits on the value of the median of the population. We perform bootstrap resampling of the 10 data values $B = 10,000$ times, computing the median each time. If the resulting bootstrap median values are arranged in ascending order and labelled from 1 to 10,000, so that the minimum value is x_1 and the maximum value is x_{10000}. Which of these points is the lower limit of a two-sided 95% CI? Which one is the upper limit (e.g., x_{10} and x_{100})?

6. **(2 points)**
If A and B are events such that $P(A) = 2/5$ and $P(B) = 5/7$, can A and B be disjoint events? If not, what is the minimum value of $P(A \cap B)$?

7. **(4 points)**
A random variable X is distributed according to $P_X(x) = C\left(\frac{2}{3}\right)^x$ for $x = 0, 1, 2, \cdots$.
(a) Find C. **(b)** If $Y = \dfrac{X}{X+1}$, what is the distribution of Y, and what are the values Y can have?

8. **(3 points)**
Probability integral transformation: Let X be a random variable with CDF $F_X(x)$. Assume that X has a continuous distribution (so that the CDF has a unique inverse). If Y is a random variable such that $Y = F_X(x)$, show that $Y \sim U[0,1]$.

9. **(1 point)**
X is a random variable drawn from an unknown continuous PDF with a finite standard deviation. What is the largest possible value of the probability that a randomly drawn value of X is at least 5 standard deviations away from the population mean, regardless of the direction of the deviation?

10. **(1 point)**
The mean of a sample of $N = 50$ points drawn from an unknown distribution is 25, with a standard deviation of 10. Construct a 95% CI for the population mean.

Please use `Python` functions to answer the following questions.

11. **(6 points)**
A scientist constructs a star's spectral energy distribution (SED) using broadband photometry measurements at 12 wavelengths from the UV through the mid-infrared.

    (a) Assuming that the SED can be fit with a blackbody model (2 parameters, so $\#\text{dof} = 12-2 = 10$), the scientist produces a fit resulting in a χ^2 value of 3.32. What is the p-value associated with this determination? Based on this value, do you accept or reject the scientist's blackbody hypothesis at a 95% confidence level?

    (b) At a 95% confidence level, what is the range of acceptable χ^2 values for these data?

12. **(13 points)**
The K-band luminosity function of carbon stars in the Large Magellanic Cloud.
This problem will use K-band photometry from the 2MASS survey for stars in the Large Magellanic Cloud that were classified as carbon-rich by Boyer et al. (2011). The data is available here in the form of a two-column comma-separated file, with the first column containing the K-band magnitude and the second column containing the uncertainties in these magnitudes. In what follows, the histogram of K-band magnitudes will be referred to as the K-band luminosity function (KLF).

    (a) Plot the KLF. Generate $N_{\text{iter}} = 1000$ realisations of this KLF and compute the 95% CI for each bin. Combine these CIs to generate a 95% point-wise confidence band for the KLF. Overlay this confidence band onto the original KLF.

    (b) The location of the peak of the carbon-star LF can place strong constraints on the efficiency of the third dredge-up process (see, *e.g.*, Marigo et al. 1999). Generate $N_{\text{iter}} = 1000$ realisations of the KLF using the magnitude uncertainties. For each realisation, find the magnitude at which the KLF peaks. Use these values to compute a 95% CI for the magnitude of the KLF peak. Plot the computed range onto the figure generated in 12a.

Warning: make sure that the bin edges/locations don't change during the multiple realisations of the KLF!