Statistics for Astronomy 2019–2020
Exam
4 November 2019 (8:30 - 11:30)

DIRECTIONS: ALLOW 3 HOURS. WRITE YOUR NAME AND STUDENT NUMBER AT THE TOP OF EVERY PAGE OF YOUR SOLUTIONS. Please explain clearly all of the steps that you used to derive a result. Please make certain that your handwriting is readable to someone besides yourself.

1. Answer the following open questions: **7 points/question**

   (a) Assume that Jeffrey's prior is used as prior probability, $\mathrm{prob}(x|I)$, for variable x. Use the transformation rule to calculate the probability distribution function of the (natural) logarithm of x, $\mathrm{prob}(\log x|I)$.

   (b) Using Cox' rules, derive the mathematical formula that describes Bayes' theorem and give the name of each term in the formula.

   (c) Describe and compare **parameter estimation** and **model comparison**. What role does the evidence play in each process?

   (d) Describe the **principle of indifference** and give a simple example of its application to determining a probability distribution function.

   (e) Given a continuous uniform distribution function $\mathrm{prob}(x) = \frac{1}{8}$ for $0 \leq x \leq 8$ and $\mathrm{prob}(x) = 0$ otherwise:

      i. Calculate $\left\langle x^3 + 1 \right\rangle$

      ii. Calculate $\langle x \rangle$ and $\mathrm{Var}(x)$, the variance of x.

   (f) Explain what the Levenberg-Marquardt algorithm is used for, and describe in words how it works.

   (g) There are several ways to discover exoplanets, including radial velocity searches. In a radial velocity search, an exoplanet is found by measuring the change of the radial velocity of its host star. Such an exoplanet is often characterized by the period P and eccentricity e of its orbit. These aren't the only parameters that need to be inferred: the model of a planet's orbit *also* includes the mean radial velocity of the star V_0, the velocity amplitude of the orbit K, the "longitude of the periastron" ω, and the fraction of the orbit prior to the start of the data taking at which the periastron occured χ. Given some data $\mathbf{D}$ (which includes the velocities of the host star measured at some Julian Dates):

      i. name and describe the process by which you could estimate the **joint probability distribution function** $\mathrm{prob}(P, e|\mathbf{D}, I)$ from the full posterior probability distribution $\mathrm{prob}(V_0, K, P, e, \omega, \chi|\mathbf{D}, I)$ determined from the orbital model; and

      ii. describe how you would determine the **best estimates** of the period P_0 and eccentricity e_0 of the orbit and the **uncertainties** on these best estimates σ_P and σ_e.

   (h) A particular data set has N values y_i $(0 \leq i < N)$ at positions x_i. Each value y_i is independent and its uncertainty follows a Gaussian distribution with the same, but unknown standard deviation. The data set is modelled by a straight line, $y = a + bx$. Given the data and the fitted values for a and b, write down the equation for the best estimate of the variance of the random uncertainty. You do not have to show a derivation in your answer.

2. Open questions involving derivations

   (a) (**16 points**) Given a data set $\{x_i : 0 \leq i < N\}$ of N values, where each value x_i is independent and drawn from the same normal (Gaussian) distribution with known (constant) standard deviation σ, use Bayes' theorem to prove that the most probable value μ_0 for the mean of the set is the average of the data. Assume a flat prior for the mean μ.

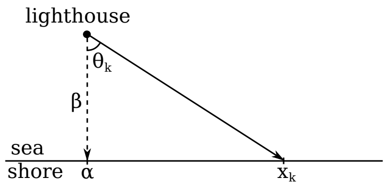

Figure 1: The Lighthouse problem

(b) (**12 points**) Consider "the lighthouse problem" (Fig. 1): A lighthouse at sea is at a position α along the coast and a distance β at sea. The shore is lined with photodetectors that record only that a flash has been recorded, but not the angle from which it came. N flashes have been recorded at positions $\{x_i : 0 \leq i < N\}$ along the coast, corresponding with azimuth directions $\{\theta_i : 0 \leq i < N\}$. The relation between a detected position and azimuth direction is given by $\beta \tan \theta_i = x_i - \alpha$. We know that the lighthouse flashes uniformly in the directions of interest: $\text{prob}(\theta_i|\alpha, \beta, I) = \frac{1}{\pi}$ for $-\frac{\pi}{2} \leq \theta_i \leq \frac{\pi}{2}$.
Calculate $\text{prob}(x_i|\alpha, \beta, I)$.
Hint: the table below contains some of the common trigonometric derivatives.

function	derivative	function	derivative
$\sin x$	$\cos x$	$\arcsin x$	$1/\sqrt{1-x^2}$
$\cos x$	$-\sin x$	$\arccos x$	$-1/\sqrt{1-x^2}$
$\tan x$	$\sec^2 x$	$\arctan x$	$1/(1+x^2)$

(c) (**6 points in total**) Three questions about the previous result:

i. What is the name of the specific distribution derived in question 2(b)?
ii. There is a more generic distribution that can also take the form of the function derived in question 2(b). What is its name?
iii. When is this generic distribution equal to the specific distribution of question 2(b)?

3. True/false questions – mark T for a true statement or F for a false statement on your exam paper: **1 point/question**

(a) You should always use least-squares to fit a straight line to data.

(b) A binomial distribution is an appropriate likelihood function for binned data when you know both the expected value and expected variance of an experiment.

(c) A χ^2 test is the appropriate test when comparing two continuous distributions or a continuous distribution with a model that predicts a continuous distribution.

(d) The distribution of the mean of a set of values drawn from a (one-dimensional) Gaussian distribution with a constant but unknown uncertainty can be calculated from the Student's t-distribution.

(e) The central limit theorem says that all probability distributions tend towards a Gaussian distribution in the large-N (many observation) limit.

(f) The best estimate of a parameter drawn from a (one-dimensional) Gaussian distribution based on data with unequal uncertainties is the weighted mean of the individual data points.

(g) A Poisson distribution is an appropriate likelihood function for binned data when you know the expected signal in each bin.

(h) The mean of the Poisson distribution is equal to square root of its variance.

(i) The Kolmogorov-Smirnov statistic is the appropriate test when comparing two distributions generated from binned data, for example when asking whether two luminosity functions have been drawn from the same distribution.

(j) There are three horses and an unknown number x of alpacas in a barn. We know that there are at most 5 alpacas in the barn ($0 \leq x \leq 5$) and the probability $\mathrm{prob}(x \mid I)$ is uniformly distributed. An unknown animal escapes. The probability that the animal escaping is an alpaca, $\mathrm{prob}(\mathrm{alpaca} \mid I)$, is given by the formula

$$\mathrm{prob}(\mathrm{alpaca} \mid I) = \frac{1}{6} \sum_{x=0}^{5} \frac{x}{3+x}.$$