Statistics for Astronomy 2020–2021
Exam
28 October 2020 (8:30 - 10:30)

DIRECTIONS: **Allow 2 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: (**10 points/question**)

   (a) Given N mutually exclusive hypotheses $\{H_i\}$ that together cover all possible options. Consider the case of calculating $\text{prob}(A \mid I)$ from $\text{prob}(A, H_i \mid I)$ using marginalization, where I is some background information.

      i. Write down Cox's product and sum rules.

      ii. What equation is implied by having N mutually exclusive hypotheses $\{H_i\}$ that together cover all options?

      iii. Use these equations to derive the mathematical formula that describes the discrete marginalization rule.

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

   (c) In the context of model comparison, describe in words what the "Ockham factor" is.

   (d) Given a continuous distribution function $\text{prob}(x) = \frac{1}{2}x$ for $0 \leq x \leq 2$ and $\text{prob}(x) = 0$ otherwise, calculate the expectation value $\langle x \rangle$ and the variance of x, $\text{Var}(x)$.

   (e) For certain parameter estimation problems, the most probable estimate of the parameters can be solved by minimizing the χ^2 function, i.e., by doing a least-squares fit.

      i. Write down the four requirements necessary for least-squares fitting to produce the most probable estimate.

      ii. Consider the case of fitting parameter a in model function $f(x) = a\sin(x)$ to a set of N measurements $\{y_i\}$ with corresponding uncertainties $\{\sigma_i\}$ and positions $\{x_i\}$, with $0 < i \leq N$. The position values $\{x_i\}$ can be considered to have absolute certainty (i.e., have no errors). Write down the χ^2 function that needs to be minimized.

   (f) Given the two-dimensional Gaussian PDF,

$$\text{prob}(x, y \mid \sigma, I) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right),$$

   and given the relationship between R, θ and x, y of

$$x = R\cos\theta \qquad \text{and} \qquad y = R\sin\theta \text{ —}$$

   that is, if x, y are Cartesian coordinates, R, θ are in polar coordinates — calculate the PDF $\text{prob}(R, \theta \mid \sigma, I)$ by transforming the PDF $\text{prob}(x, y \mid \sigma, I)$.

→ **See next page for questions 2 and 3**

2. (**30 points**) Given a data set of measured values $\{y_i\}$ (with $0 < i \leq N$) with corresponding positions $\{x_i\}$. Start from Bayes' theorem and derive the formula for the most probable value a_0 and its uncertainty σ_a for the parameter a in the equation $y_i = a\, x_i$. The random errors of $\{y_i\}$ are independent and drawn from a normal (Gaussian) distribution. All values y_i have the same uncertainty, given by σ_y. The position values $\{x_i\}$ can be considered to have absolute certainty (i.e., have no errors). Assume a flat prior for a.

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

   (a) The Metropolis-Hastings algorithm is a method for obtaining a sequence of random samples, to step through the parameter space of a PDF.

   (b) For a position parameter x for which no prior knowledge is available, one should use the following prior: $\log \mathrm{prob}(x) = \text{constant}$.

   (c) Bootstrapping is a method to create multiple data sets from a single data set.

   (d) The central limit theory states that, when comparing two models that fit the data equally well, the simplest model is more likely to be correct.

   (e) The Poisson distribution is given by:

   $$\mathrm{prob}(k) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

   (f) The mean of an experiment with two outcomes that is repeated multiple times is distributed according to the Cauchy distribution.

   (g) The mean of an independently normally distributed data set with known uncertainties follows a normal distribution.

   (h) Student's t distribution is an appropriate likelihood function for binned data when you know the expected signal in each bin.

   (i) MCMC (Markov chain Monte Carlo) methods can be used to analyse the posterior of complex, non-linear models.

   (j) Assume that for dinner, a single alpaca eats between a half and one (uniformly distributed) kilograms of grass and hay on a day. Consider X, the combined weight of the food for a large group of N alpacas on a single day. X will tend towards a normal distribution as N increases.