Statistics & Data Analysis Problem Set

S. Biller

1. A friend asks you to guess a secret number between 0 and 1000. You will have five guesses and, after, each one, you will be told if your guess is too high, too low, or correct. Show that a binary search, where each guess is chosen to divide the available phase space into equal portions, has the expectation of ending up closest to the correct number.

2. Assuming the birthdays of the population are equally distributed through-out the year, how many people, taken at random, need to gather together in a room to give a greater than 75 percent chance of at least two of them sharing a common birthday?

3. Consider the Poisson probability distribution:

$$P(n|\mu) = \frac{\mu^n e^{-\mu}}{n!}$$

(a) Show that this is a well-behaved probability in that: $\sum_{n=o}^{\infty} P(n|\mu) = 1$ and $\int_0^{\infty} P(n|\mu)d\mu = 1$.

(b) Show that the mean value of the probability distribution $\langle n \rangle = \mu$.

(c) A famous story about the use of Poisson statistics is attributed to Ladislaus Bortkiewicz, who was commissioned by the Czar of Russia to determine whether the rate at which Prussian Army soldiers were being kicked to death by horses was due to chance or the wrath of god. He published his results in 1898 after studying 14 calvary corps over the period from 1875-1894:

Number of deaths/year/corps	Observed Frequency
0	109
1	65
2	22
3	3
4	1
>4	0
Total	200

Calculate the mean number of deaths per corps. Compare the observed frequency with that predicted by the Poisson distribution with this calculated mean.

How well does this match expectation from real, random data?

4. This question is about a continuous probability distribution known as the exponential distribution. Let x be a continuous random variable that can take any value $x \geq 0$. It is said to be exponentially distributed if it takes values between x and $x + dx$ with probability:

$$P(x)dx = Ae^{-x/\lambda}$$

where λ and A are constants.

(a) Find the value of A that makes P(x) a well-defined continuous probability distribution so that $\int_0^\infty P(x)dx = 1$.

(b) Show that the mean value of the probability distribution is
$\langle x \rangle = \int_0^\infty xP(x)dx = \lambda$.

(c) Find the variance of this probability distribution.

Both the exponential distribution and the Poisson distribution are used to describe similar processes, but for the exponential distribution x is the actual time between successive radioactive decays, successive molecular collisions, or successive horse-kicking incidents (rather than x being simply the number of such events in a specified interval).

5. Analyse the following using Bayes' Theorem: Mrs. Trellis (from NorthWales) has 2 children, born 3 years apart. What is the probability that Mrs. Trellis has a daughter under each of the following conditions:

a) Her son likes Marmite.
b) At least one of her children is a son.
c) At least one of her children is a son who likes Marmite.
d) At least one of her children is a son who likes Marmite, assuming that this survey was commissioned by Marmite.

6. Suppose you are looking for evidence of a rare particle produced following a cosmic ray interaction in your detector. The particle decays in the characteristic exponential fashion with a mean lifetime of τ, and you aim to look for a signal produced by the by-products of this process. However, there is also a constant level of background interactions due to the decays of radioactive isotopes in your detector that can mimic the signal, so you plan to look for an excess of events above this background within a given time period following the cosmic ray event. What is the optimal time window duration to choose so as to maximise the sensitivity of your search?

7. Imagine fitting a set of N data points (x_i, y_i with standard errors σ_i) with a straight line of the form $y = mx + b$. By minimising the appropriate χ^2 function, show that the best fit values for slope m and intercept b are:

$$m = \frac{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i y_i}{\sigma_i^2} - \sum \frac{x_i}{\sigma_i^2} \sum \frac{y_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \sum \left(\frac{x_i}{\sigma_i^2}\right)^2}$$

$$b = \frac{\sum \frac{y_i}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \sum \frac{x_i y_i}{\sigma_i^2} \sum \frac{x_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \sum \left(\frac{x_i}{\sigma_i^2}\right)^2}$$

8. Generally, background estimates have uncertainties associated with them that should be taken into account when estimating significance levels. Consider a simple counting experiment divided into two different measurements where a certain amount of "off-source" data is first taken to provide an estimate of background levels in order to test for a possible excess of events in a second "on-source" data set. Let E_{off} represent the off-source exposure (basically the product of the detector acceptance and counting time), during which N events are observed. Let E_{on} represent the on-source exposure, during which n events are observed. Define $\alpha = E_{on}/E_{off}$ and take B to represent the "true" mean of the background counts for the off-source exposure while S represents the "true" mean of excess signal events seen on-source.

(a) Construct a likelihood function for the two observations under hypothesis $H0$: that $S = 0$, so both data sets should be treated equally as only having background. Construct another likelihood function under $H1$: that $S > 0$ and the data sets should thus be treated differently.

(b) Given the observations, what values of S and B would maximise the likelihoods in each case?

(c) Write an expression for the log of the likelihood ratio, $\mathcal{L}_R$, here defined as $\mathcal{L}(H1)/\mathcal{L}(H0)$. Show that:

$$-\log \mathcal{L}_R(max) = N \log \left[(1+\alpha)\frac{N}{N+n}\right] + n \log \left[\left(\frac{1+\alpha}{\alpha}\right)\frac{n}{N+n}\right]$$

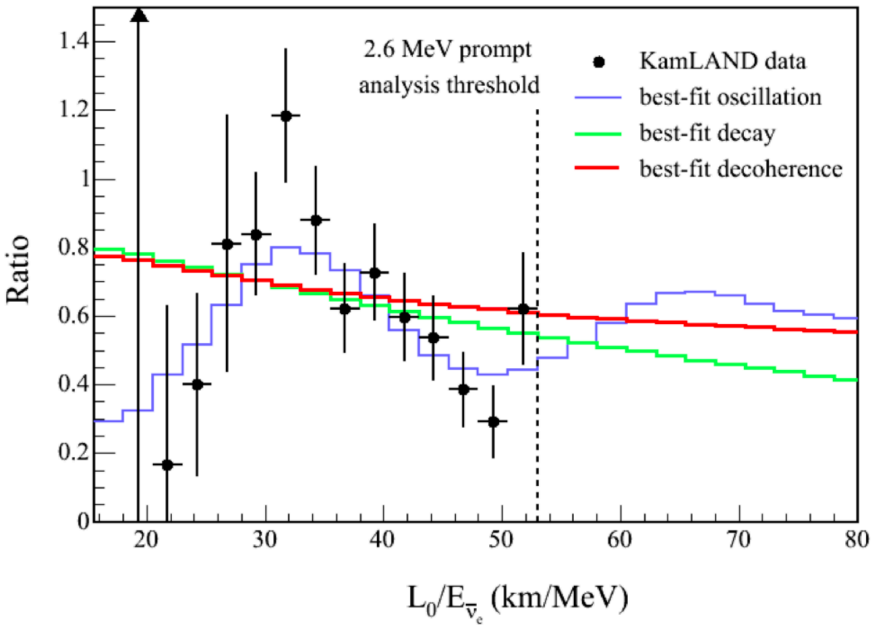
(d) How is $-\log \mathcal{L}_R$ distributed? From this, write an expression for the equivalent number of standard deviations pertaining to the significance of an observed excess in the on-source data relative to expectations from the off-source data.

9. A model for Earth's atmosphere predicts that the temperature T at a height h above Earth's surface varies as $T = T_0 - Lh$, where T_0 is the surface temperature and L is known as the adiabatic lapse rate. On a certain day, measurements give $T_0 = 295 \pm 2K$ and $L = (6.5 \pm 0.5) \times 10^{-3} Km^{-1}$. Determine the expectation value of T and its experimental uncertainty for a weather balloon at a height $h = 2800 \pm 200m$.

10. The KamLAND experiment has studied the oscillation of anti-electron neutrinos from nuclear reactors around Japan using a large, liquid scintillation detector. The fundamental parameters that define the nature of the oscillation are the mixing angle, θ_{12}, and the difference in squared masses between the states, Δm^2_{12}. The probability for an electron (anti)neutrino to "survive"without oscillating to another state before reaching the detector is:

$$P_{surv} \simeq 1 - sin^2(2\theta_{12})sin^2(\alpha \Delta m^2_{12} L/E)$$

where L is the distance travelled by the neutrino in km, E is it's energy in GeV and α has a value of 1267 MeV-radians per eV^2-km. Below is a plot from their 2004 paper showing the ratio of observed events to predictions for no-oscillations as a function of L/E, assuming an average distance of reactors from the detector of 180 km. The data is compared to the best fit model for oscillations (blue histogram) as well as with two non-standard models of neutrino decay (lower green line) and decoherence (upper red line). For all models, you may assume that 3 parameters were varied in the fit.



(a) Does the oscillation model provide an acceptable fit to the data?

(b) Can either the decay or decoherence models be ruled out based on this data?

(c) Identify the data point with the most significant deviation from the oscillation model. What is the chance probability of observing a deviation at least this large if the oscillation model is correct?

(d) Based on the error bar sizes, estimate the statistical uncertainty in the overall fit normalisation to the survival probability.

(e) The average baseline (*i.e.* distance weighted by the expected neutrino flux) of nuclear reactors from KamLAND is $\sim$180 km. Assume that the systematic uncertainties in the average reactor distance and the neutrino energy determination are about 2%. The distribution of data points might be used to approximate the statistical uncertainty in determining the position of various features. Use this and the information above to estimate the value of Δm^2_{12}, the individual contributions to its uncertainty and the total, combined uncertainty on the parameter determination. State any assumptions you make.