DATA 221 - Homework 1 - Spring 2023

Trimble

Due: Friday 2023-03-31 11:59pm

Reading: MacKay Chapter 2 and 3

1. Henry Newson reported a series of measurements of the decay rates of ^{17}F produced by deuteron bombardment of oxygen, producing the following measurements of decay rate as a function of time since the accelerator was turned off. (Henry W. Newson. "The Radioactivity Induced in Oxygen by Deuteron Bombardment." Phys. Rev. 48, 790 (1935) doi:10.1103/PhysRev.48.790)

time (min)	Decay rate (arbitrary)
0.161	87.1
0.578	69.7
1.113	51.5
1.584	38.3
2.226	25.3
3.061	15.8
4.324	7.6
6.229	2.5

Decays of this source occur at a rate that decays exponentially with time:

$$r(t) = Ae^{\frac{-\ln(2)t}{t_{1/2}}} + B$$

We are interested in estimating the decay constant $t_{1/2}$ from observations of the rate.

The decay rate numbers originate from counts of decays in a fixed time period, but we can't learn exactly how many decays were detected because the count data ("rate") has been modified by an unstated affine transform.

   a. Find the unweighted least-squares fit for the values of A, B, and $t_{1/2}$. (Don't use this number!)
   b. Weight the sum-squared error terms by (1/sqrt(rate)) and report the fitted values of A, B, and $t_{1/2}$.

2. That was a curve-fitting problem where the measured quantities were rates. Consider a different sort of problem, instead of summing over bins, what if we sum over the data points?

   "Unstable particles are emitted from a source and decay at a distance x, a real number that has an exponential probability distribution with characteristic length λ." In other words, x is exponentially distributed. Let us imagine a magical counter that can measure decays near $x = 0$ and decays infinitely far away from the source. The counter observes six events, $x_n = 1.5, 2, 3, 4, 5, 12$, and is certain other events occurred.

   The phrase above "exponential probability distribution with characterisic length λ" is code for "The distances are exponentially distributed according to $p(x) = \frac{1}{\lambda}e^{\frac{x}{\lambda}}$ for $x > 0$".

a) Using a prior density that is proportional to $\frac{d\lambda}{\lambda}$ (which is the appropriate prior density for a scale parameter), find and plot the posterior density for λ.

b) Give a 95% confidence interval for λ.

c) Estimate the mean of the posterior density (by numerical integration).

Hint: this problem uses a different parameterization for the exponential distribution than the first; Q1 uses half-life; Q2 uses mean-life, which is ln(2) longer. (Simplifed version of Problem 3.3 from Mackay, p 47.)

3. A die is selected at random from three twenty-faced dice on which the symbols 1–10 are written with non-uniform frequency as follows:

Die	1	2	3	4	5	6	7	8	9	10
A	6	4	3	2	1	1	1	1	1	0
B	3	3	2	2	2	2	2	2	1	1
E	2	2	2	2	2	2	2	2	2	2

A randomly chosen die from A, B, or E is rolled 7 times, with the following outcomes: 5, 3, 9, 3, 8, 4, 7. Then, a randomly chosen die from all three is rolled 8 times, with the following outcomes: 5, 3, 9, 3, 8, 4, 7, 10.

a) What are the probabilities that the die is die A, B, or E after the first seven rolls?

b) What are the probabilities that the die is die A, B, or E after rolling the 10?

c) What are the consequences of the zero probability for die A to return a 10 ?

d) How many rolls on average would you need to establish 99:1 confidence between B and E? (Hint: there is a theoretical answer (sums over things) but you could get an answer by simulation.)

(Exercise 3.1 from MacKay p.47)

4. Given a distribution of bigrams in English text, the distribution of the initial letter given the final and that of the final letter given the initial are different.

a) Reproduce the three figures in Fig 2.1 and 2.3 in MacKay (Hinton diagrams or heatmaps are fine) using the novella Carmilla by Joseph Sheridan Le Fanu (`https://www.gutenberg.org/files/10007/10007-0.txt`).

b) Count the one-letter tokens.

c) Split the text into two-letter tokens and count them.

d) Make a two-dimensional visualization of the bigram frequency, row-marginal, and column-marginal probabilities. (Hint: one of your visualizations should show that Q is always followed by U)