1. **Markov Chain Monte Carlo (AST 518 Students only)**

The Bayesian analysis of the linear fit program from HW 2 was hampered by the need to do a 3-dimensional integral for each value of m and b, essentially a 5-d grid calculation. While this was conceptually more straightforward, there is a more efficient method which uses Markov Chain Monte Carlo calculations. Here, we will modify the previous problem (HW 2, problem 3) to carry this out. It will have the additional advantage of giving us an estimate for the parameters of the bad data points, and the confidence intervals for the fit parameters. You will want to modify your program along the following lines:

i. Define a function to calculate the maximum likelihood function, given a vector of the 5 parameters $(m, b, P_b, V_b, Y_b)$. Make sure $ML=0$ for nonphysical values of the parameters such as $P_b<0$. Define a set of starting values for your parameters and calculate $ML$.

ii. Define a function that changes the vector of parameters to new values in some defined way. You can cycle through each parameter, or randomly choose one to change. You want this to be a random variation with appropriate variation for the problem. It is this function where the art of MCMC is found.

iii. If the new set of parameters has $ML_{new}>ML_{old}$, accept it. If $ML_{new}<ML_{old}$ accept it with a probability of $ML_{new}/ML_{old}$.

iv. Repeat a large number of times ($1e4-1e5$). Each time a new set of parameters is accepted, store it for analysis at the end of the program.

Once you have a working program, you should be able to answer the following:

a. Evaluate the Markov chain to determine how many points are needed to converge to a solution. Eliminate these points from further analysis. Try different starting points and ways of generating new steps to optimize this. What worked/ didn’t work?

b. Create histograms of the 5 fit parameters showing their range of variations in the MCMC program.

c. Estimate the 95% confidence intervals for the slope and intercept parameters using the Markov chain data directly. That is, calculate what range about the mean encompasses 95% of the data points. Why is this a better description of the parameter uncertainty than quoting a 2 sigma value for the parameters? Do the two values differ for this situation?
2. **Malmquist Bias Simulation.**

We want to understand how brightness limits affect the use of supernovae observations in the measurement of the Hubble constant, $H_0$. To do so, we will use Monte Carlo computations to generate a set of simulated data. For this simulation, we make the assumption that we are only using a class of supernovae that all have similar intrinsic brightness, so that their apparent flux can be used to measure how far away they are. If their Doppler shift is also measured, the measurements can be used to calculate the expansion rate of the universe, or Hubble's constant.

A subtle effect occurs in this measurement, called Malmquist bias, that can affect the result. The effect is caused by the range of apparent brightness for supernova. Supernova in our simulation have an absolute magnitude of $M = -19$. Assume the supernova have a scatter about their absolute magnitude of approximately 1 magnitude. Survey telescopes will detect objects as faint as $m = 20$. The limiting magnitude corresponds to a distance modulus $(m-M = -5 \log_{10}(d/10))$ of 39, which suggests that supernova can be seen as far away as 1000 Mpc.

Assume supernova are formed uniformly throughout a sphere with radius $r = 2000$ Mpc. In the data generation part of the simulation, assume that each supernova is receding at a rate $v = H_0 \cdot d$, where $H_0$ is 72 km/s / Mpc and $d$ is the distance in Mpc.

a. Create a Monte Carlo program to generate a user-selectable number of randomly placed supernovae within this volume. Have the program generate the true distances, $d$, to the supernovae. Using the program, calculate the mean distance for a set of supernova. You will need to use the transformation method to generate properly distributed distances. You can also calculate the mean distance analytically. Make sure the result agrees with your Monte Carlo answer.

b. Now assume each of the supernovae has a brightness governed by $M = -19 + G(1)$

where $G(1)$ is a random number with Gaussian distribution and standard deviation of one magnitude. Calculate the apparent magnitude of each star, using the distance generated in a. If $m > 20$, assume the object is too faint to detect, and reject it from the sample.

Use the remaining points, to simulate the observations by generating a velocity for each one, using Hubble's law. Generate an observed distance ($d'$) by using its apparent magnitude and Hubble's law, with the assumption that the supernova has an absolute magnitude ($M$) of 19. Create a simulated observation sample of 100 supernovae. Plot these with observed distance on the $x$-axis and velocity on the $y$-axis and compare it to the original sample. Explain the effect of the observing limit on the resulting sample.

c. Calculate $H_0$ by averaging $v/d'$ for all your detected points. Discuss the level of the bias. How could you account for it in the observations?