Continual Assessment Number 2

1. In our local patch of the galaxy, the value of m is observed to be roughly 0.5, but this has only been measured for stars that are more massive than those found in the modeller's simulations. What can the researcher say about the mean and standard deviation of m given the outcome of her simulations?

After her 3 simulations, she had no multiples and therefore m for her simulations was 0. This example can be defined with a binomial function (either a multiple system or a single system), therefore we can create a likelihood distribution of m.

We can write an expression for the binomial distribution which can be written as :-

$$P(\nu|N,m)=\binom{N}{\nu}m^{\nu}(1-m)^{N-\nu} \quad (1)$$

By using multiple different values for m, we can find the distribution which can be seen in Figure 1.

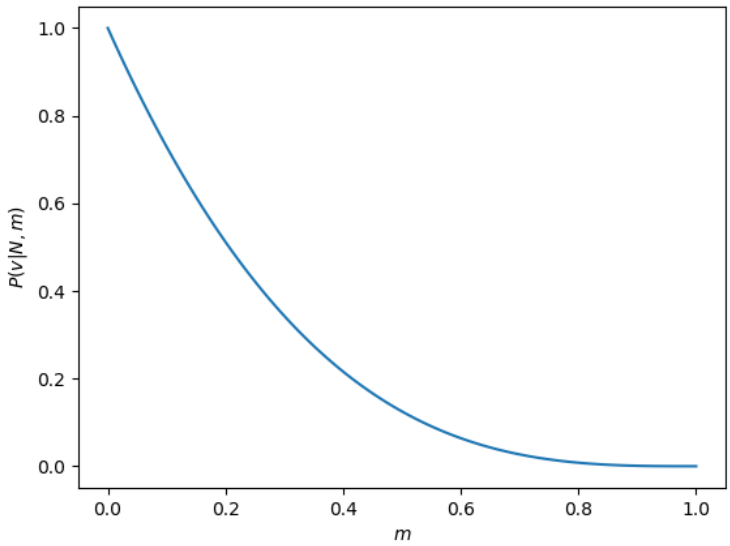

Figure 1: Graph of the likelihood distribution of the mean, m, given by a binomial distribution.

Using Bayes Theorem, we can estimate the mean and standard deviation of m given her simulations, where the expression for Bayes theorem is :-

$$P(\theta|x)=\frac{P(x|\theta)P(\theta)}{\int_{\theta_{min}}^{\theta_{max}} P(x|\theta)P(\theta)}d\theta \quad (2)$$

Where θ is m in our case, and x is our data from the simulations.

P(θ) is the prior, which can be found using some prior information such as a previous study. We can use the information given in the question for our prior using some sort of distribution. From the question it states that the value of 0.5 for m was measured only for stars that are more massive than those found in the simulations.
We can create a prior distribution using either use a flat prior or a beta distribution depending on the prior information.

A flat prior is flat distribution where for all values of m the probability is 1 :-

$$P(m)=1 \qquad for\ 0<m<1$$

A beta distribution is a more general form of the Binomial distribution.

In our case a beta distribution can be written as :-

$$P(m|a,b)=\frac{m^{(a-1)}(1-m)^{(b-1)}}{B(a,b)} \tag{3}$$

Where B(a,b) is simply a normalisation factor.

Given the result of 0.5 for m in previous study, we must choose suitable values for a and b such that the mean of the prior distribution is 0.5, the mean of the beta distribution is given by :-

$$\hat{m}_B=\frac{a}{a+b} \tag{4}$$

For m to be 0.5, a and b have to be equal, but we still have a problem of determining the value of a/b. A large a/b will give a well defined distribution and a small a/b will give a broad distribution.

The standard deviation of a beta distribution is also given in our notes and is expressed as :-

$$\sigma_B=\sqrt{\frac{\hat{m}_B(1-\hat{m}_B)}{a+b+1}} \tag{5}$$

Since the prior information suggests that since m=0.5 for stars only more massive than those in the 3 simulations, one can assume a broad beta distribution as a prior because the prior information suggests that there have been single systems in the previous study (otherwise m would be 1 if no single systems were found) and those single systems have all been more massive than the results of the simulation suggesting that the prior information could have no relevance to the simulation results whatsoever. However, I think that there could be some relevance in the prior information as the question states “In our local patch of the galaxy” which may suggest that the patch of galaxy that determined the value of 0.5 could’ve been small and only a few stars were calculated and could mean that although its not very accurate it could be useful to include since if we extending the patch of sky could produce results of single systems with masses similar to the results of the simulation, or by repeating the simulations we might get systems that are more massive.

I chose a beta distribution for my prior with values of a=2 and b=2. The value of 2 was chosen because the information of for the prior is vague and so the distribution should reflect that, and the value of 2 produces a very broad beta distribution which can be seen in Figure 2.

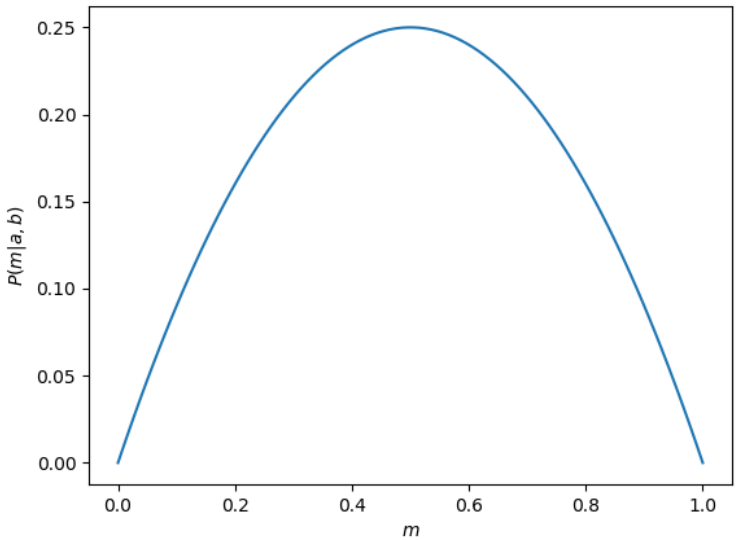

Figure 2: Graph of the prior distribution of m, given by a beta distribution where a=b=2.

Using both distributions we can then get a posterior distribution using bayes theorem and equation 2, this is done by simply multiplying both distributions together and normalizing by dividing by the sum of the two distributions multiplied together to get a distribution as seen in Figure 3.

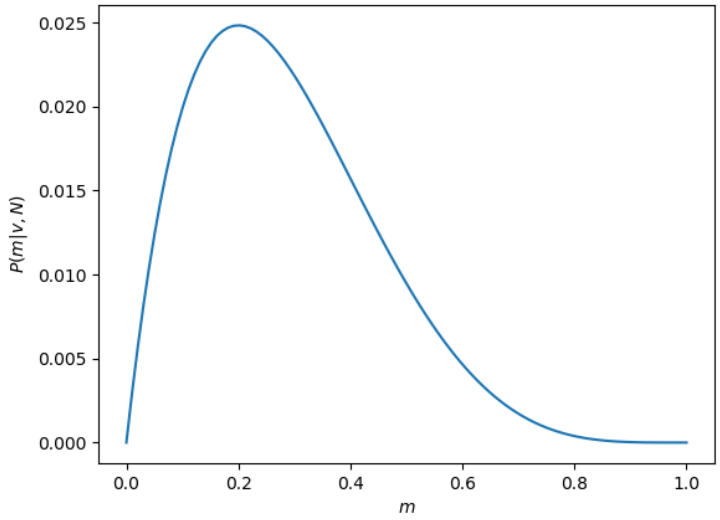

Figure 3: Graph of the posterior distribution of the mean.

The posterior distribution can be expressed in terms of a beta distribution and in turn we can find the mean and the standard deviation of the distribution using equations 4 and 5 by using the fact that a ≡ ν + a and b ≡ N - ν + b :-

$$\hat{m} = \frac{\nu + a}{N + a + b} \tag{6}$$

Where $\hat{m}$ is the mean of the posterior distribution.

$$\sigma = \sqrt{\frac{\hat{m}(1-\hat{m})}{N + a + b + 1}} \tag{7}$$

Where σ is the standard deviation of the distribution.

The mean of the posterior was found to be 0.2857 and the standard deviation of the posterior was found to be 0.1597. Therefore the researcher can say that the mean and standard deviation of m given the outcome of her simulations is 0.286 and 0.160 respectively.

2. The remnants are taken to be radiocarbon-dated, and found to have the following ages (in years):

2141.22 1781.15 1523.37 1816.90 1932.29 1541.21 720.782 1026.22 1687.55 2460.59

A previous study, in a neighbouring valley, recorded the eggs to be 1200 years old, following a normal distribution with standard deviation of 300 years.

A. Write down an expression for the posterior distribution for the mean age of the newly found eggs in the region.

Given that our data can be taken to follow a normal distribution, we can write the likelihood of the mean of the data to be a normal distribution :-

$$P(\hat{X}|\theta) = N\left(\theta, \frac{\sigma^2}{n}\right) \tag{9}$$

Where $\hat{X}$ is the mean of the data, σ is the standard deviation of the distribution and θ is some unknown parameter that describes the mean of the data (in years) and is used to find the most probable mean of the data (through the mean of the posterior distribution)

And using a similar normal distribution for the prior :-

$$P(\theta)=N(\mu_{0,}\sigma^2{}_0) \tag{10}$$

Where μ_0 is the mean of theta for the prior (previous data) and σ_0 is the associated error.

From equations 9 and 10, we can use Bayes theorem to find the most probable theta :-

$$P(\theta|\hat{X})=\frac{P(\hat{X}|\theta)\,P(\theta)}{\int_{\theta_{min}}^{\theta_{max}} P(\hat{X}|\theta)\,P(\theta)}\,d\theta \tag{11}$$

Using equations 9, 10 and 11 we can find the posterior distribution in terms of the new data and the previous data.

$$P(\theta|\hat{X})=\left(\frac{1}{\sigma/\sqrt{n}\sqrt{2\pi}}\right)\exp\left(\frac{-(\hat{X}-\theta)^2}{2\sigma^2/n}\right)\left(\frac{1}{\sigma_0\sqrt{2\pi}}\right)\exp\left(\frac{-(\theta-\mu_0)^2}{2\sigma^2{}_0}\right) \tag{12}$$

Since were multiplying two normal distributions together, we would expect another normal distribution for the posterior distribution. Therefore if we consider another normal distribution that describes theta in terms of the mean and standard deviation of the theta, we can then find the mean and standard deviation of the posterior in terms of the values of the new data and the prior data.

$$P(\theta)=\left(\frac{1}{\hat{\sigma}\sqrt{2\pi}}\right)\exp\left[\frac{-(\theta-\hat{\theta})^2}{2\hat{\sigma}^2}\right] \tag{13}$$

Equating equations 12 and 13 can yield the mean and standard deviation.

$$\hat{\sigma}=\left(\frac{1}{\sigma^2/n}+\frac{1}{\sigma^2{}_0}\right)^{-1/2} \tag{14}$$

$$\hat{\theta}=\left(\frac{\sigma^2{}_0}{\sigma^2{}_0+\sigma^2/n}\right)\hat{X}+\left(\frac{\sigma^2/n}{\sigma^2{}_0+\sigma^2/n}\right)\mu_0 \tag{15}$$

The posterior distribution was plotted using equation 12, Where $\hat{X}=1663.1$, $\sigma/\sqrt{n}=151.8$, $\mu_0=1200$ and $\sigma_0=300$ and can be seen in Figure 4.

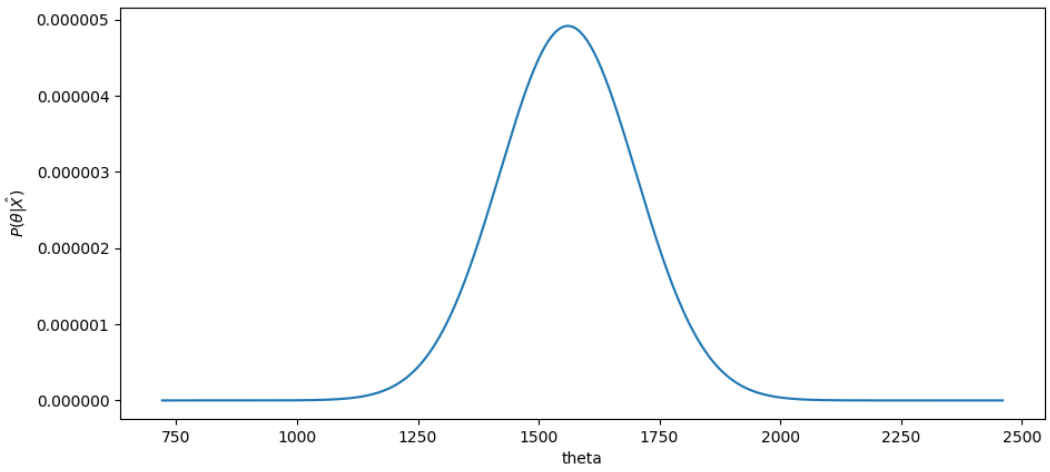

Figure 4: Analytical graph of the posterior distribution.

From the analytical distribution and equations 14 and 15, the mean and standard deviation of the posterior distribution was found to be 1568.75 years and 135.42 years respectively.

B. Write your own MCMC algorithm and use it to create a posterior distribution for the mean age of the eggs.

Following the instructions given for the metropolis walk on pages 49 and 50 in our notes I created an MCMC that targets the posterior distribution.

Using equation 12 as the target distribution (the P(θ) in our notes), I generated histogram of theta which should in theory replicate the shape of the posterior distribution this can bee seen in Figure 5.

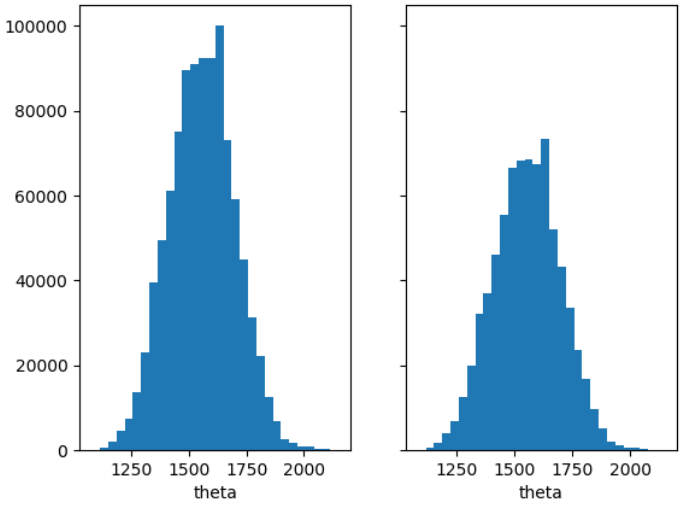

Figure 5: On the left we have a histogram of the MCMC that was run for a total of 1000000 steps, on the right we have pulled the last 750000 values and ignored the first 250000 from the MCMC to take into account the 'burn in' time.

From Figure 5, we can see that the MCMC histogram resembles the shape of the posterior that can be seen in Figure 4. We can also see that the Condensed sample has an ever so slightly better shape. The method and code used to find the MCMC posterior distribution can be found below.
A plot of the probability distribution can also be created by inputting the values of theta back into the posterior distribution function and a plot can be made as seen in Figure 6. Compared the the analytical distribution in Figure 4, we can see that the MCMC algorithms distribution in Figure 6 is very close to the analytical distribution.

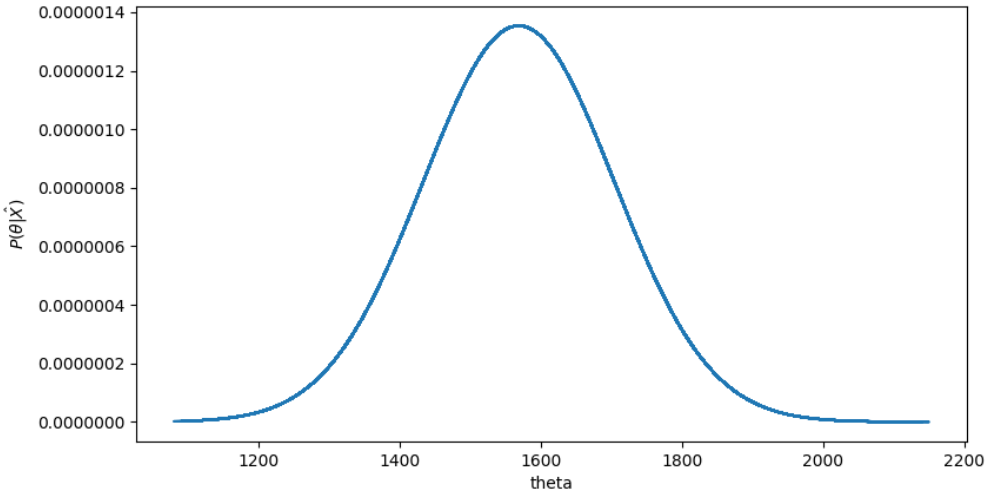

Figure 6: A posterior distribution of the mean created

C. Use your MCMC results to compute the mean and standard deviation for the posterior.

Using the numpy module in python we can compute the mean and standard deviation of the MCMC histogram, using np.mean() and np.std().
The mean and standard deviation using the MCMC algorithm was found to be 1553.9 years and 144.4 years respectively, which are quite close to the analytical solutions of the posterior distribution of 1568.75 years and 135.42 years respectively.

D. Use MCMC to work out the evidence term, and compare to the analytic expression.

The evidence term can be found using equation 5.21 in our notes which states that :-

$$\frac{1}{p(D)} \approx \frac{1}{N} \sum_{\theta_i \sim p(\theta_i|D)}^{N} \frac{h(\theta_i)}{p(D|\theta_i)\,p(\theta_i)} \tag{16}$$

Where p(D) is our evidence term, $p(D|\theta_i)p(\theta_i)$ is our posterior distribution, $h(\theta_i)$ is a normal distribution with a similar shape to the posterior distribution and θ_i are values pulled from the posterior distribution using an MCMC algorithm.

To compute the evidence term I used the condensed sample of the MCMC as my values of θ_i because these values were pulled from the posterior distribution. The mean and standard deviation of the normal distribution used for $h(\theta_i)$ was the mean and standard deviation computed from the MCMC algorithm which was 1544.6 years and 150.7 years respectively.

The value of p(D) was found to be 0.000449, which was very accurate compared to the analytical solution of the evidence term which was found to be 0.000459.

(The source for my analytical solution was wolfram alpha : http://www.wolframalpha.com/input/?i=integrate+(1%2F((300*151.76)*2*pi))*e%5E((-(x-1200)%5E2)%2F(2*300%5E2))+*+e%5E((-(x-1663.1282)%5E2)%2F(2*151.76%5E2))++from+500+to+2500) and can be seen in Figure 7.

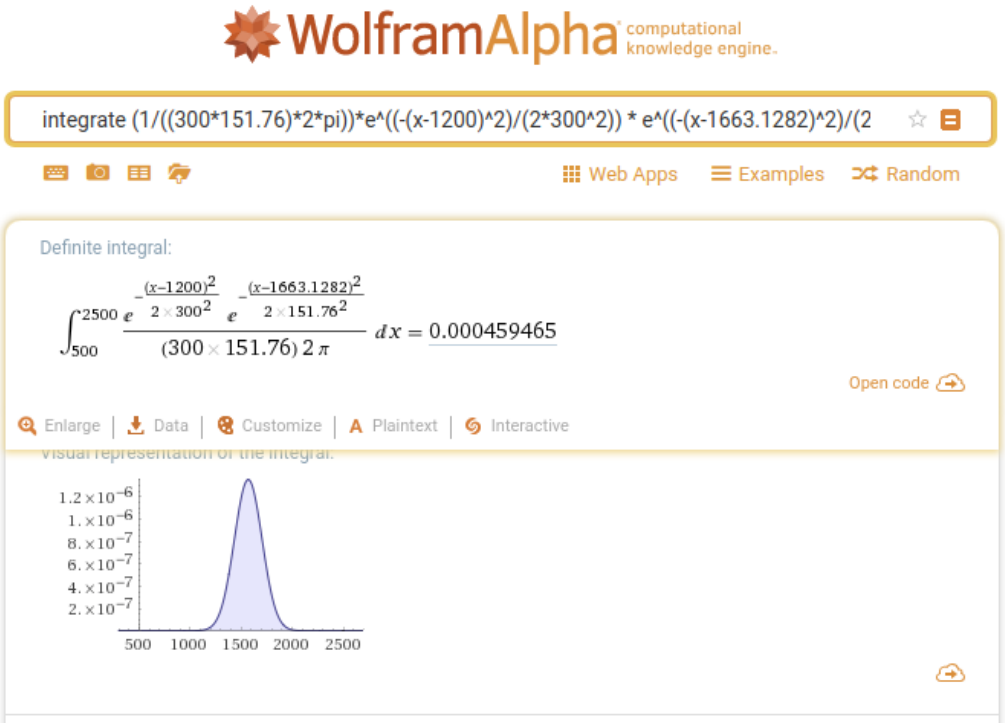

Figure 7: Screenshot of the analytical solution for the evidence term

The analytical solution for the evidence term was found using the fact that :-

$$p(D)=\int p(D|\theta)\,p(\theta)\,d\theta \tag{17}$$

And so I simply integrated equation 12 with the limits between 500 and 2500 as below 500 and above 2500 the probability is approximately zero.

3. Using the Bayesian approach to hypothesis-testing, what are the odds that the tomatoes are organic?

Let us consider two different models , M1 is the model for the Organic tomatoes and M2 is the model for the Genetically modified tomatoes.

We can find the odds in favour of M1 over M2 or visa versa using the ratio of the two model posteriors:-

$$\frac{p(M_1|D)}{p(M_2|D)}=\frac{p(D|M_1)\,p(M_1)}{p(D|M_2)\,p(M_2)} \tag{18}$$

We can therefore simply use the data and equation 18 to get our odds of the tomato in question is organic.

The tomato was found to be 1.2 cm, therefore D = 1.2 and from the prior information we can see that P(GM) = 0.7 and P(O) = 0.3, where GM is the Genetically modified model and O is the Organic model.

Given that the size-distributions of the samples are normally distributed, we can use a normal distribution as our likelihood for both GM and organic.

The likelihood of the tomato being organic (using a mean of 2.2 cm and a standard deviation of 0.75 cm) is simply :-

$$p(1.2|O)=\left(\frac{1}{0.75\sqrt{2\pi}}\right)\exp\left[\frac{-(1.2-2.2)^2}{2(0.75)^2}\right]=0.0656 \tag{19}$$

The likelihood of the tomato being GM is a bit more complicated as there are two types of GM tomatoes, one being normal (with a mean of 3 cm and standard deviation of 0.2 cm) and the other being cherry (with a mean of 1 cm and standard deviation of 0.2 cm). As there is no information about the proportions of the two types, I assumed that there was equal probabilities of getting cherry and normal. The likelihood of getting a GM tomato is simply adding both probabilities as seen below :-

$$p(1.2|GM)=0.5\,x\left(\frac{1}{0.2\sqrt{2\pi}}\right)\exp\left[\frac{-(1.2-3)^2}{2(0.2)^2}\right]+0.5\,x\left(\frac{1}{0.2\sqrt{2\pi}}\right)\exp\left[\frac{-(1.2-1)^2}{2(0.2)^2}\right]=0.423 \tag{20}$$

Therefore the odds that the tomatoes are organic is simply :-

$$\frac{p(O|1.2)}{p(GM|1.2)}=\frac{p(1.2|O)\,p(O)}{p(1.2|GM)\,p(GM)}=\frac{(0.0656)(0.3)}{(0.423)(0.7)}=0.155 \tag{21}$$

Therefore the odds that the tomatoes are organic is 0.155:1.