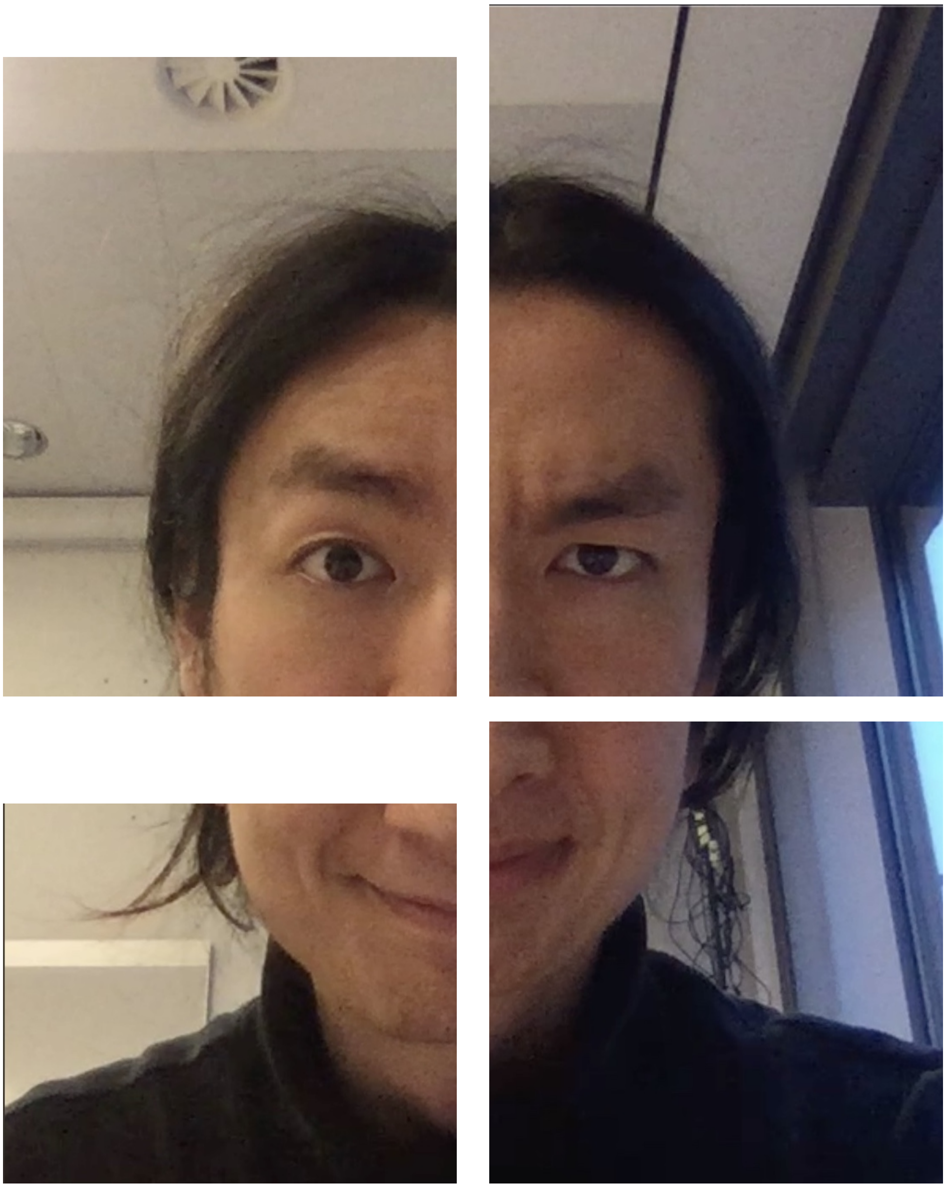
Total: 3+2.5+4 = 9.5

Problem Set #2

Advanced Methods in Applied Statistics



Cyan Yong Ho Jo

28th February 2024

EX1

Normalisation

The energy distribution of neutrinos emitted by an evaporating black hole illustrates the distribution of neutrino populations across the entire spectrum of energies. This distribution indicates the frequency of neutrinos within each energy band. Given that the mass has a significantly higher order of magnitude than temperature, performing calculations in terms of temperature is more advantageous than using mass. For example, in the provided case, the corresponding temperature is between $10-50K$, while the masses are around $10^{11}g$. Exploring the primordial energy distributions for each black hole reveals that the profile of each distribution resembles the Rayleigh distribution which is shown in Fig1.

To normalise these primordial distributions, I first needed to define the range of the distribution, as calculating an infinite number of samples is impractical. Secondly, I had to determine the area under the distribution within the defined domain. I selected an energy range of 1 to $400GeV$, based on a visual comparison of the cases. The normalisation factor is calculated using a Monte Carlo method with a scale of 100,000. The area of each distribution, along with the PDF curves, is illustrated in Fig2.

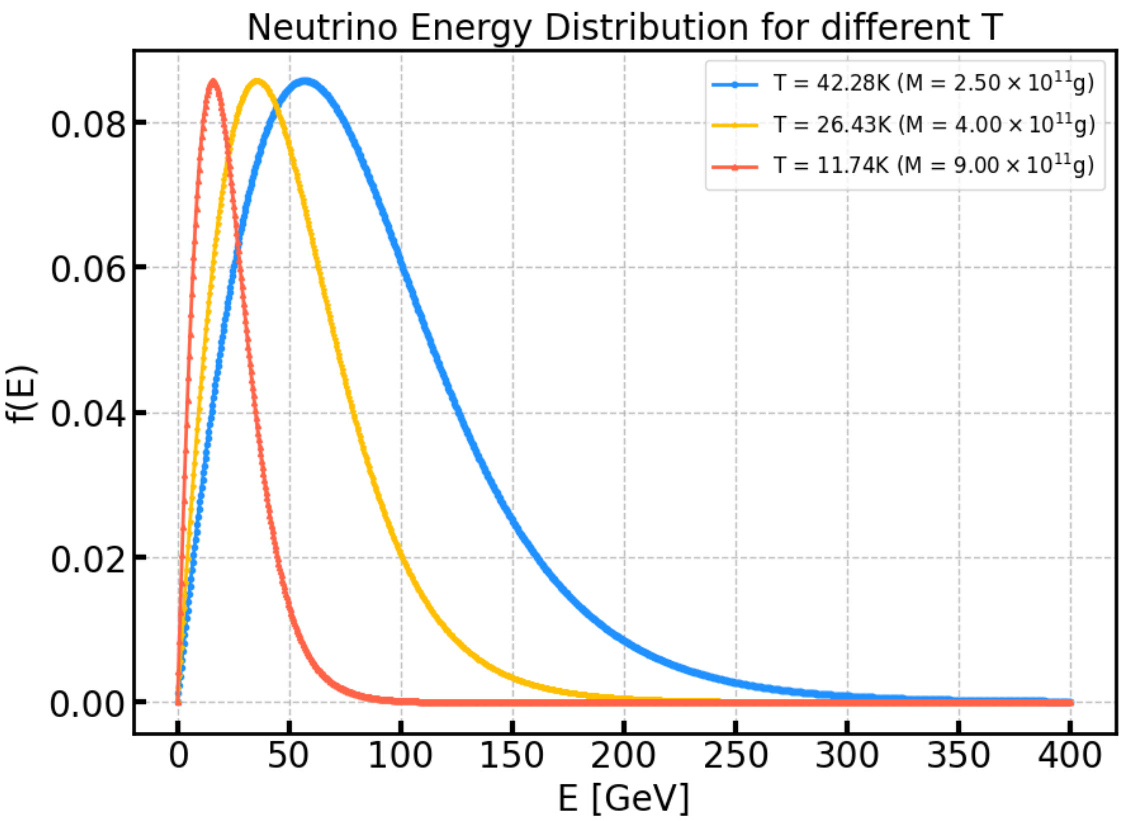
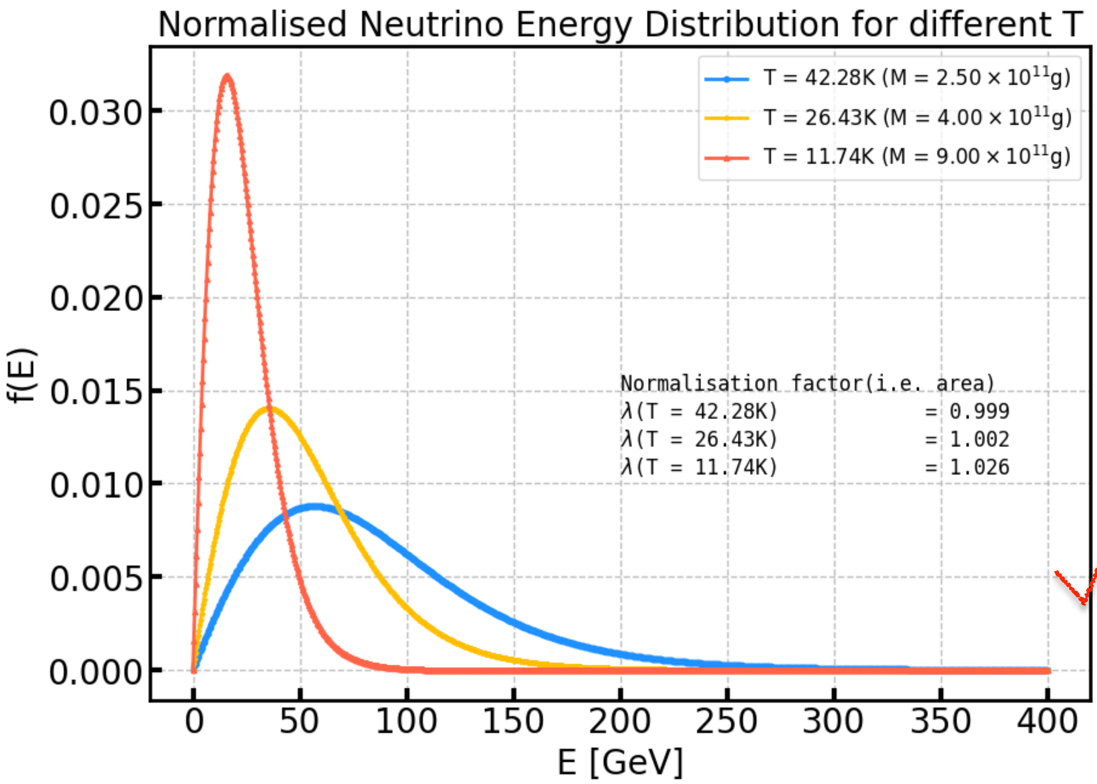


Figure 1: The primordial energy distribution of neutrinos emitted by an evaporating black hole for the black holes of given masses. The distribution is not normalised, So I use this to set a proper energy range.



0.5/0.5
Good

Figure 2: The figure shows the PDFs of neutrino energy distribution. The area or the normalisation factor of each case is shown as well in the blank area.

Histogram

For the histogram, I chose the truncated value of $\sqrt{N_{data}}$ for the number of bins which happened to be exactly 50. The histogram alone is shown in Figure3. I need to scale the PDFs I previously got to compare the histogram and the distributions in the same dimension. I introduced a new distribution which is scaled by a scaling factor λ.

$$\lambda = N_{data} \times (width)_{bin}$$

Then λ corresponds to the area beneath the scaled distribution within the same range. The scaled distributions and the histogram are shown together in Fig4. The plot hints that the mass of the black hole lies between 4.0×10^{11} and 9.0×10^{11}. My intuition urges me to insist that $M_{BH} \sim 6.0 \times 10^{11}$.

In principle, since you only have graphs for 4e11 g and 9e11 g, the best you can do would be to guess that the BH mass is around 4-9 e11 g (Without plotting more figures, you may not know exactly how (and by how much) the graphs would change as the BH mass increases).

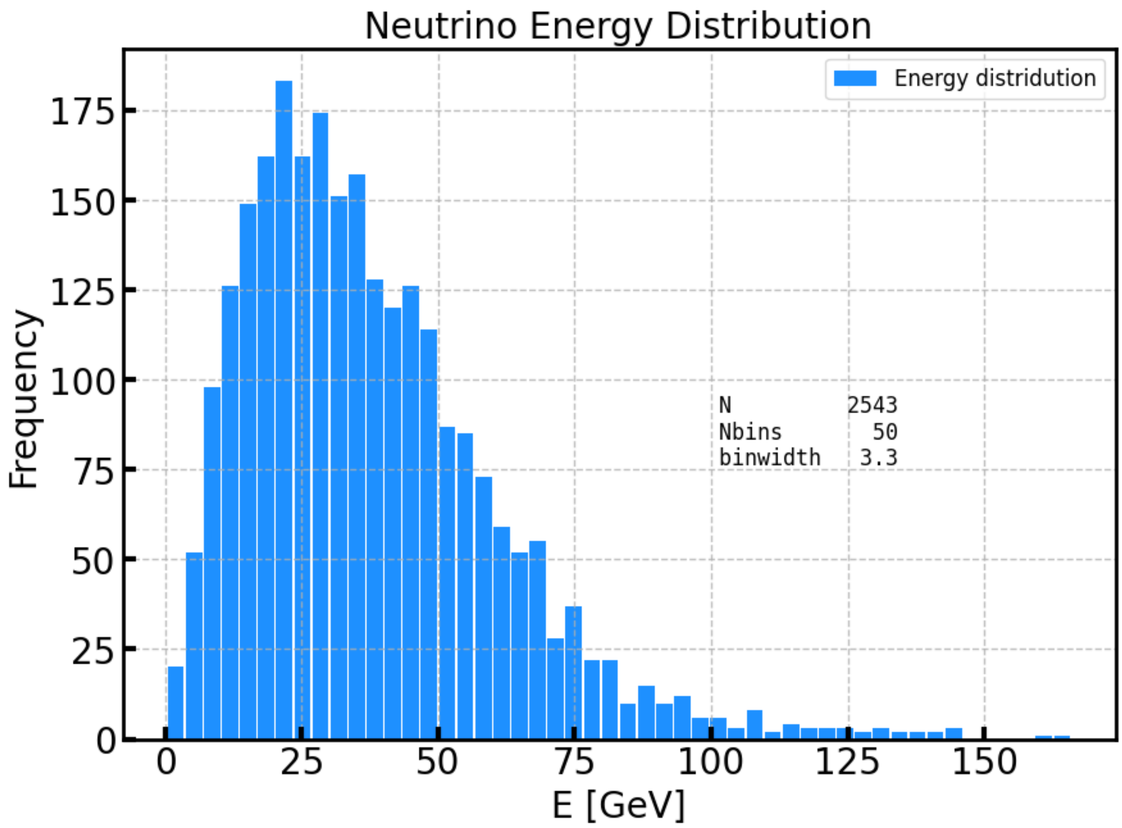
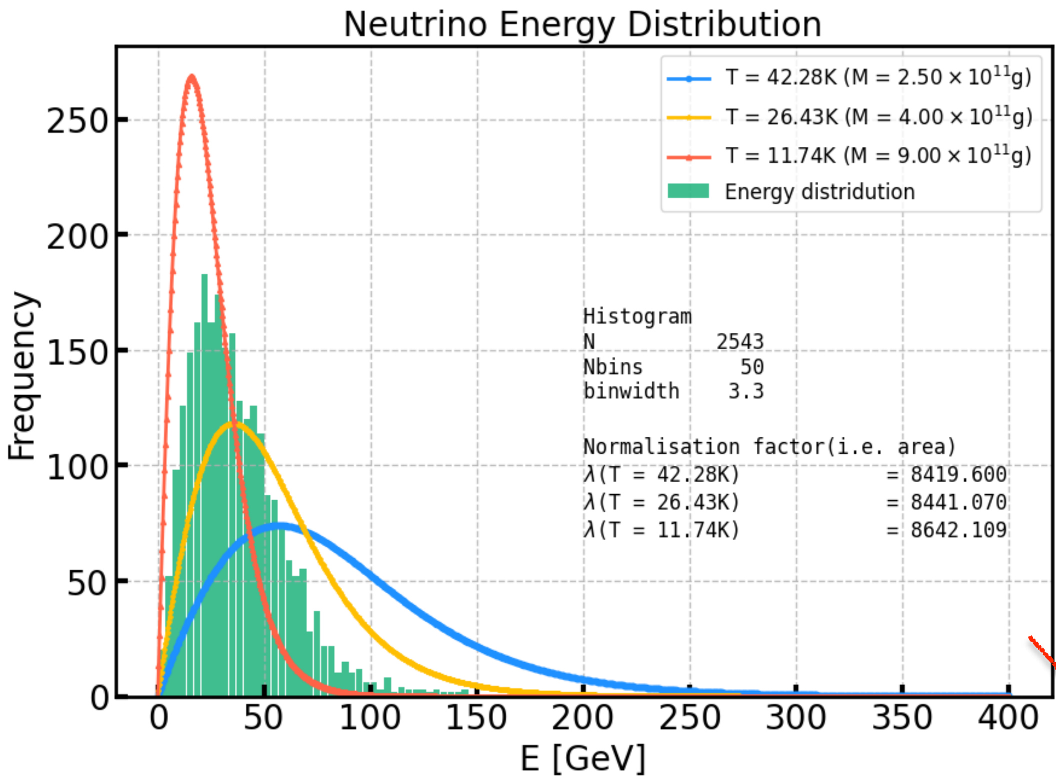


Figure 3: The histogram of the data. The number of data, the width of each bin, and the number of bins are shown in a blank area of the figure.



0.5/0.5

Figure 4: The scaled distributions plotted with the histogram in Fig3. The areas in the blank area correspond to the scaling factors or normalisation factors for each black hole case. They are not identical as the areas of PDFs are different.

1D Raster Scan

The calculation of ln(likelihood) value can be explicitly summarised as a schematic diagram in Fig**??**. In my case, I chose B for the likelihood value. The 1D raster scan is shown in Fig6. The search for the extreme value of ln(likelihood) starts at T $= 11K$ and ends at T $= 27K$. 1000 temperature values are tested. The result says the most likely values for the temperature and mass of the target black hole are $18.18K$ and $5.8 \times 10^{11} g$.

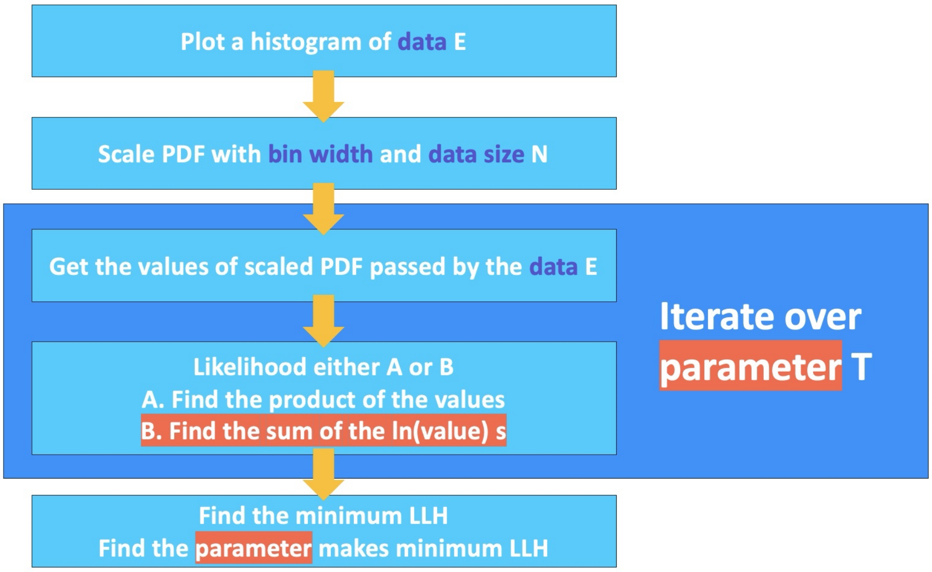
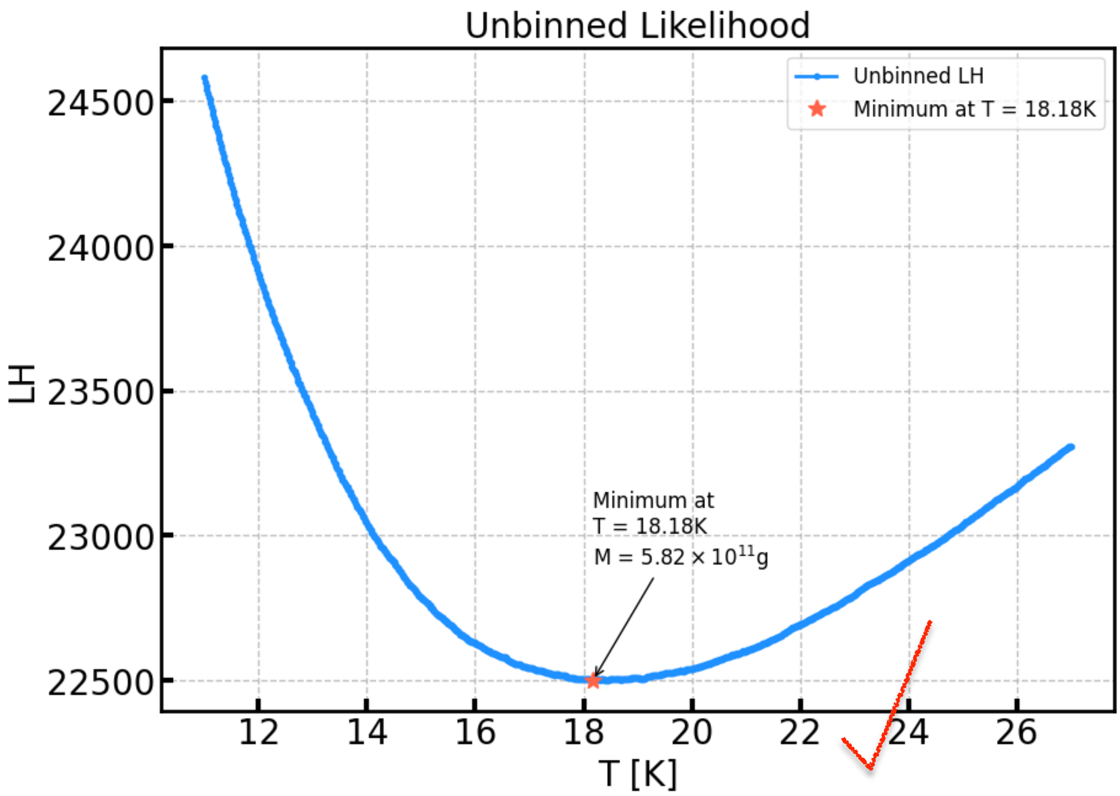


Figure 5: The visualisation of the process for calculating the likelihood values for a range of temperatures.



0.5/1

Figure 6: Raster scan of the ln(likelihood) values for the temperatures from $11K$ to $27K$. The number of steps, the minimum ln(likelihood), the temperature at the minimum ln(likelihood), and its corresponding mass are displayed in the figure.

You are probably plotting the “negative” log likelihood here.

You also need to report the likelihoods of the given masses (2.5, 4, 9).

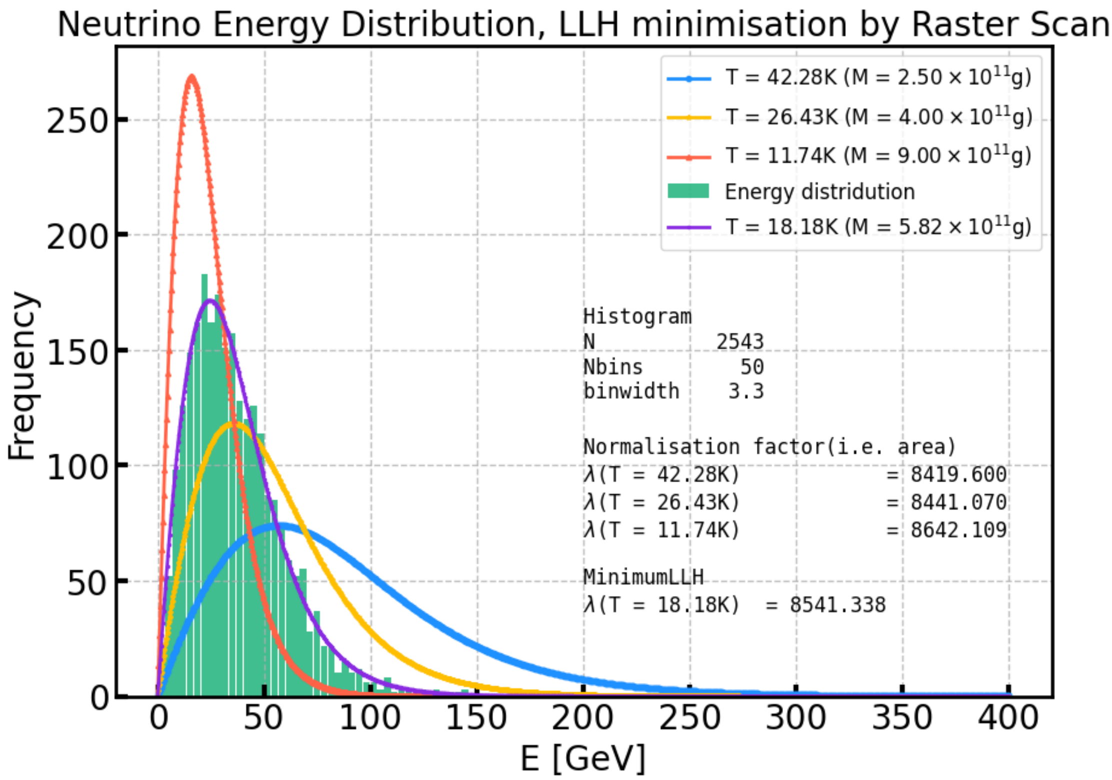


Figure 7: The histogram and scaled distributions, including the scaled distribution of black hole temperature (or mass) at the minimum likelihood, were searched using a raster scan

Minuit

The global minimum of the likelihood function was sought utilizing library functions available in *Python*. In particular, *Minuit*, implemented within the *iminuit* package, was employed for this purpose. It was observed that the default configuration of Minuit consistently failed to ascertain a minimum until the search interval was meticulously adjusted to span from T $= 16K$ to T $= 20K$, at which point the algorithm successfully identified the global minimum. The temperature and mass of the black hole, corresponding to the minimum of the negative ln likelihood, were determined to be T $= 18.58K$ and $5.69 \times 10^{11} g$, respectively.

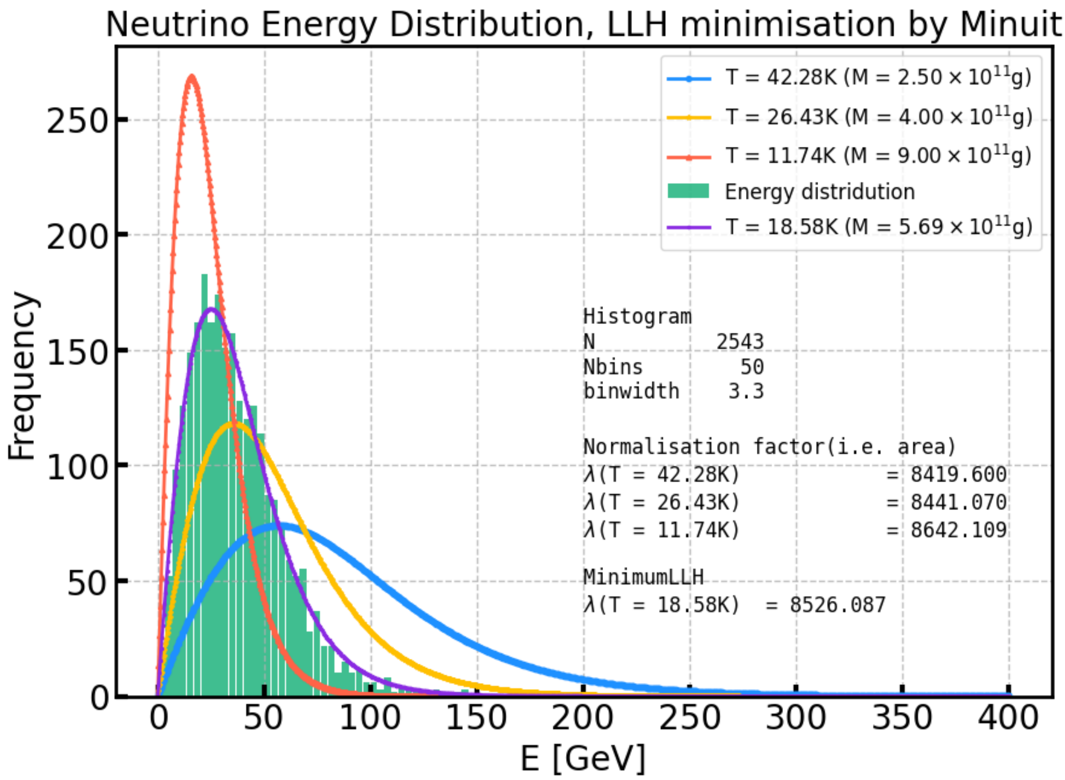


Figure 8: The histogram and scaled distributions, including the scaled distribution of black hole temperature (or mass) at the minimum likelihood, were searched using *Minuit*.

Discrepancy

There exists a discrepancy between the extreme value of the likelihood obtained through raster scanning and that identified by a library-based minimizer. This discrepancy may be attributed to variations in the algorithm and step size employed by the raster scan method compared to those utilized by Minuit.

Uncertainty

Since all the calculations were done in the temperature dimension, I need to propagate the uncertainty to get the uncertainty in the mass dimension.

$$C = 1.057 \times 10^{13}$$
$$T = \frac{C}{M}$$
$$M = \frac{C}{T}$$
$$\Delta T = -\frac{M^2}{C}\Delta M$$
$$\Delta M = M^2\frac{\Delta T}{C}$$

In the raster scan, the black hole temperature, T_{BH} was found to be $18.18K$ with a step size of $0.02K$. This can be considered to be the uncertainty for raster scan likelihood. Consequently, the true value of T_{BH} would fall within the range of $18.16K$ to $18.20K$. On the other hand, the uncertainty in likelihood using the minimiser is provided by Minuit itself, which in my case was $0.18K$. The uncertainty in the mass is calculated with the error propagation formula and the relation with the temperature. The result is shown in Table I and TableII

. height	T (K)	M (g)
Raster scan	18.18	5.82×10^{11}
Minuit	18.58	5.69×10^{11}

Table I: The result of temperature and mass determined by raster scan and Minuit minimiser.

	ΔT (K)	ΔM (g)
Raster scan	0.02	5.12×10^{8}
Minuit	0.18	5.13×10^{9}

Table II: Uncertainties in temperature and mass determined by raster scan and Minuit minimizer.

To give a single value of the black hole's temperature and mass, I calculate the weighted mean and weighted standard deviation.

$$\bar{T} = \frac{(18.18 \times (0.02)^2) + (18.58 \times (0.18)^2)}{(0.02)^2 + (0.18)^2} \approx 18.44\,\mathrm{K}$$

$$\Delta T = \sqrt{\frac{1}{(0.02)^2 + (0.18)^2}} \approx 0.02\,\mathrm{K}$$

$$\bar{M} \approx 5.81 \times 10^{11}\,\mathrm{g}$$

$$\Delta M \approx 5.10 \times 10^{8}\,\mathrm{g}$$

1.5/1.5

Problem 1 Total: 0.5+0.5+0.5+1.5 = 3

EX2

The probability density function for a gamma distribution with shape parameter $\alpha = 2$ and rate parameter $\beta = 2$ is normalised by definition, ensuring that the integral of the PDF over its domain equals one. As the rate parameter β increases, with α held constant, the distribution narrows and becomes more sharply peaked around zero. However, the area under the curve remains one, as this is a fundamental property of probability density functions.

PDFs

The posterior distribution is obtained by normalising the product of the likelihood function and the prior distribution. The normalised PDFs for the prior, likelihood, and posterior are depicted in Fig 9.

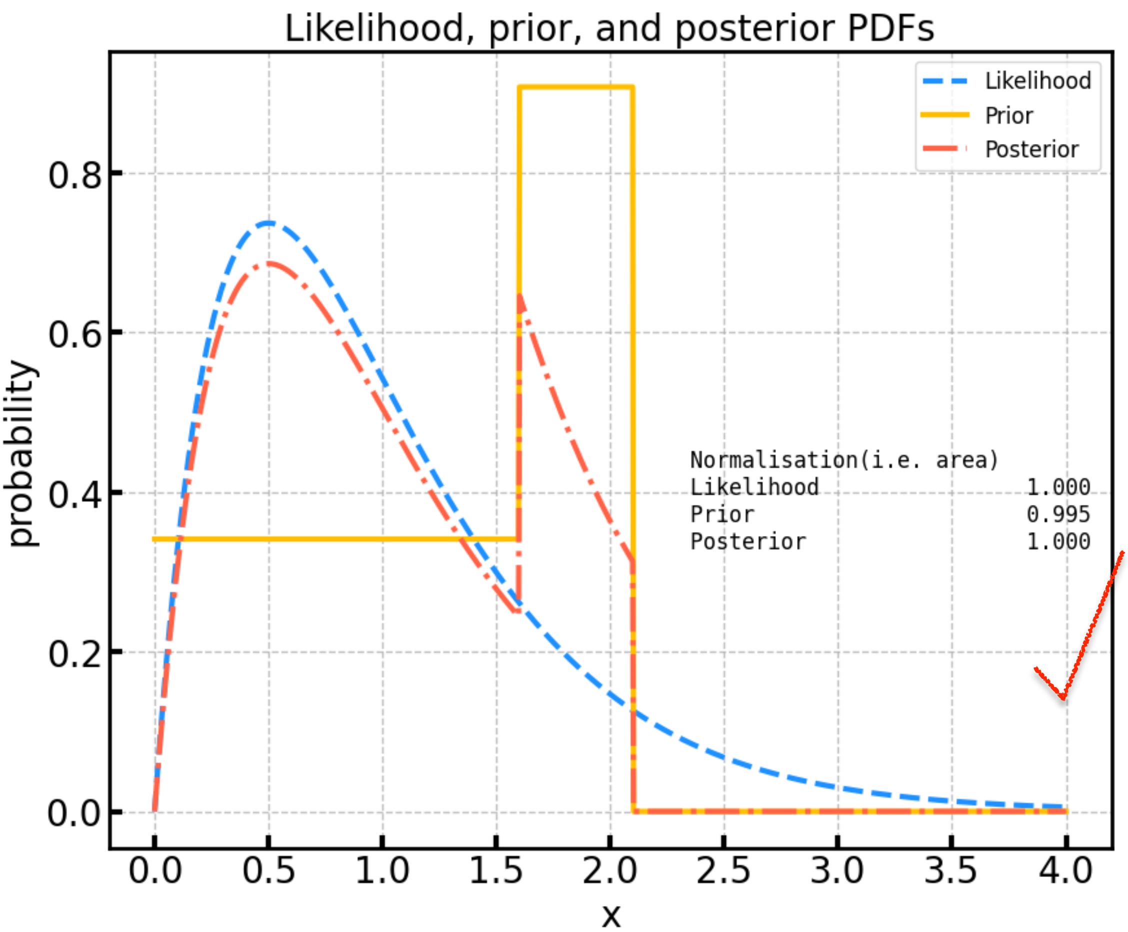


Figure 9: Normalised PDFs of the prior, the likelihood, and the posterior distributions are presented. Each distribution has been normalized to ensure that the area under its curve equals one, with these areas being illustrated in the figure as well.

Most likely point

As the posterior profile possesses two distinctive local maxima, I investigated two points: For the first region, defined by $0.3 < x < 0.7$ a local maximum was identified at $x = 0.501$. At this point, the value of the posterior probability density function is 0.686 indicating the likelihood of x being at this value within the given region. In the second region, defined by $1.5 < x < 2.0$, another local maximum was found at $x = 1.602$. The posterior PDF value at this point is 0.646. The uncertainty associated with these maxima is quantified as 0.004 originating from the step size of the x values I discretely created. I shall pick the expectation value to be the most likely value of the PDF. The expected or mean values are calculated with different step numbers in TableIII. If the most likely value is passed to the posterior distribution the returned value is 0.508 which is quite close to the location of the first local(and global) maximum.

step number	Mean	$\sum_i P(x_i)$
100	0.990	24.744
1,000	0.995	250.930
10,000	0.995	2511.139
100,000	0.995	25113.213

Table III: Summary of Posterior PDF Calculations

EX3

To tackle this problem, I first constructed a rectangular map and a circular island in m scale. I build a data structure for each crab and handle each crab object along the evolution of time.

The map

A crab on the island can travel up to 200 metres daily. Upon reaching the island's perimeter, namely the seashore, it halts. Distance travelled should be measured geometrically. Figure10 shows the configuration. Note that x takes only absolute value as the second term may exceed the first term depending on the configuration. In such a case the shaded triangle is acute instead of being obtuse. The resulting evolution of a single crab is shown in Fig11.

$$\theta = arccos(\frac{d_1^2 + d_2^2 - l^2}{2d_1 d_2})$$
$$A = \frac{1}{2} d_1 d_2 sin\theta$$
$$A = \frac{1}{2} lh$$
$$h = \frac{2A}{l}$$
$$x = \|\sqrt(d_1^2 - h^2) - \sqrt{R^2 - h^2}\|$$

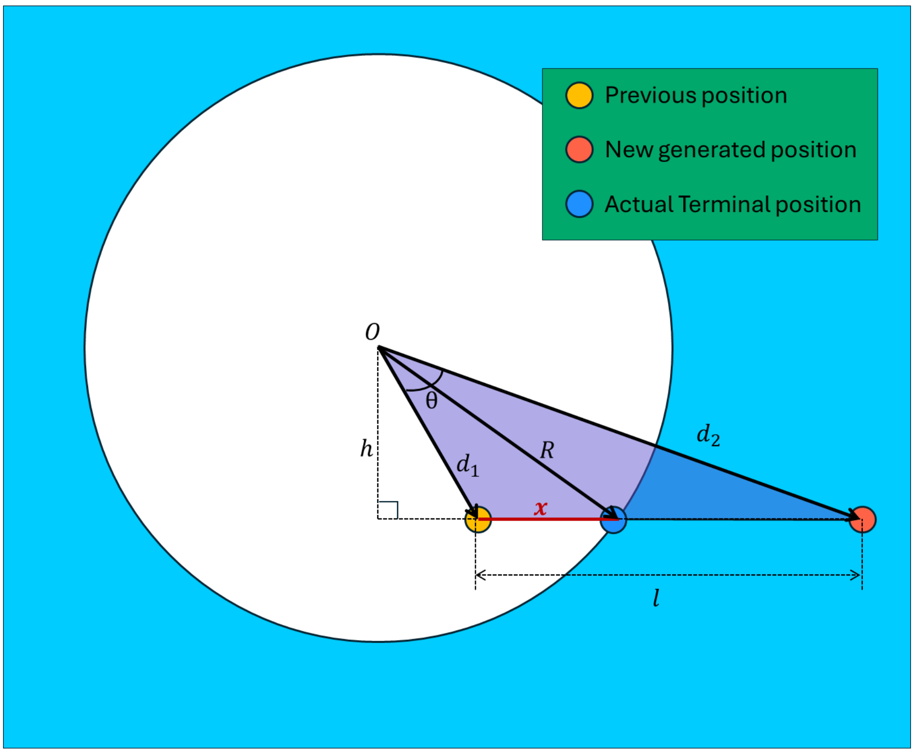
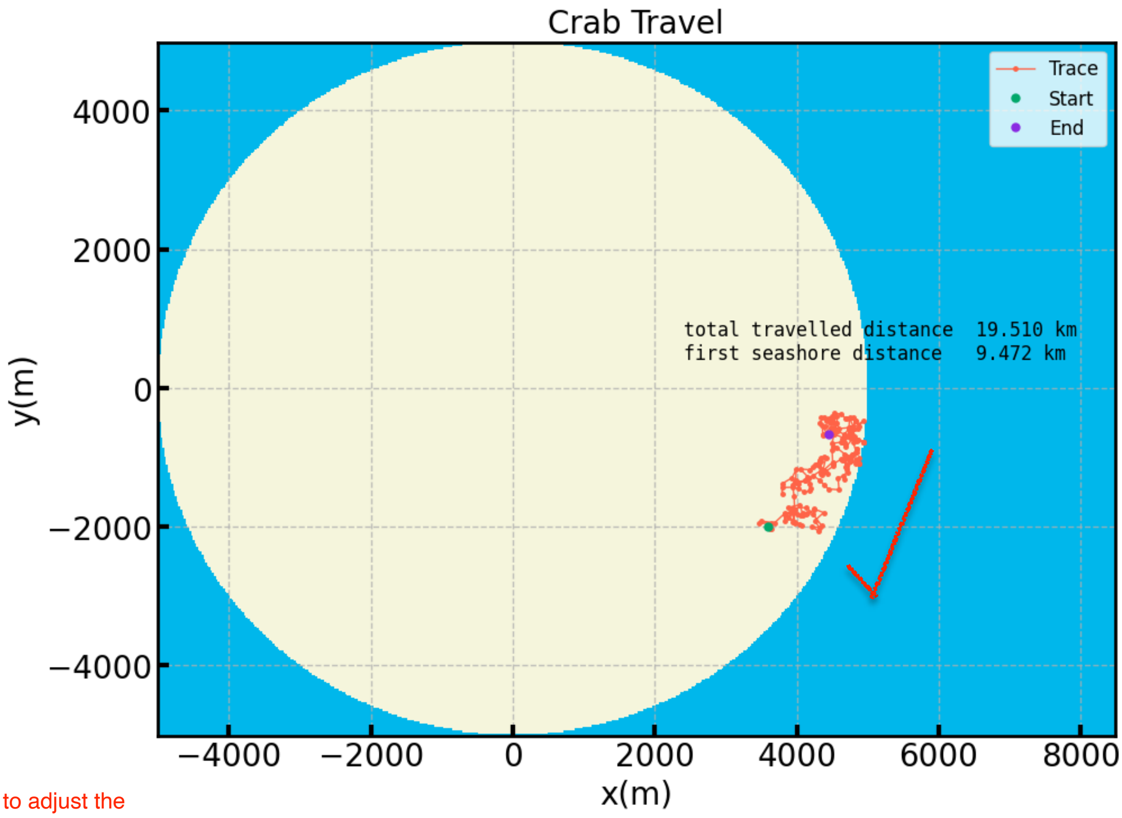


Figure 10: The distance of a crab that trying to travel over the sea.



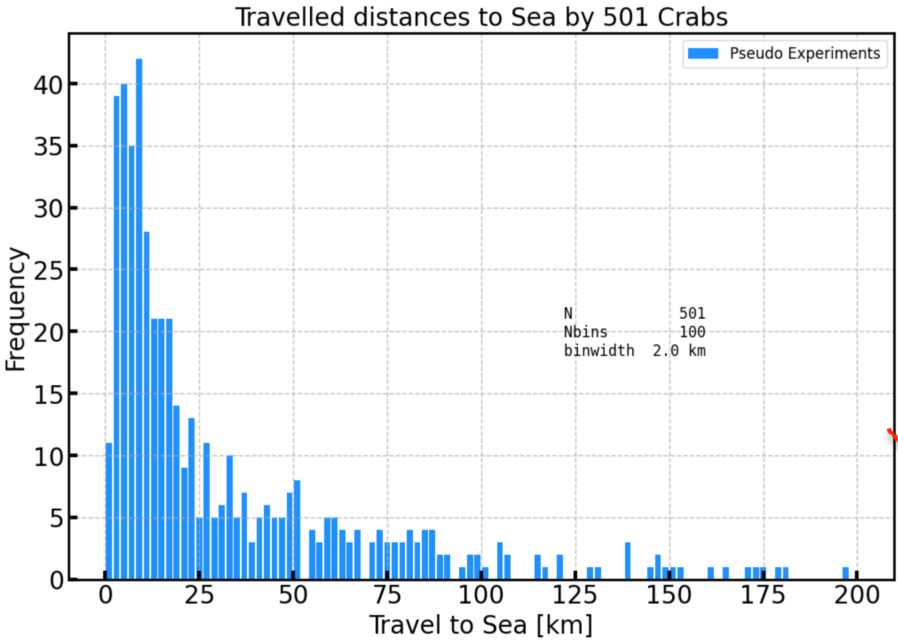
You might want to adjust the plotting region to focus on only nearby region surveyed by the crab. Given the large size of the points, sometimes it is hard to keep track of the individual day movement of the crab in your current visualization.

Figure 11: History of a crab moving on the island for 200 days.

0.5/0.5

Total distance before a crab reaches the seashore

Once it arrives at the seashore, an internal flag of a crab object stores boolean information on whether the crab has already tried to move into the sea. After 501 iterations, I got the result shown in Fig12.

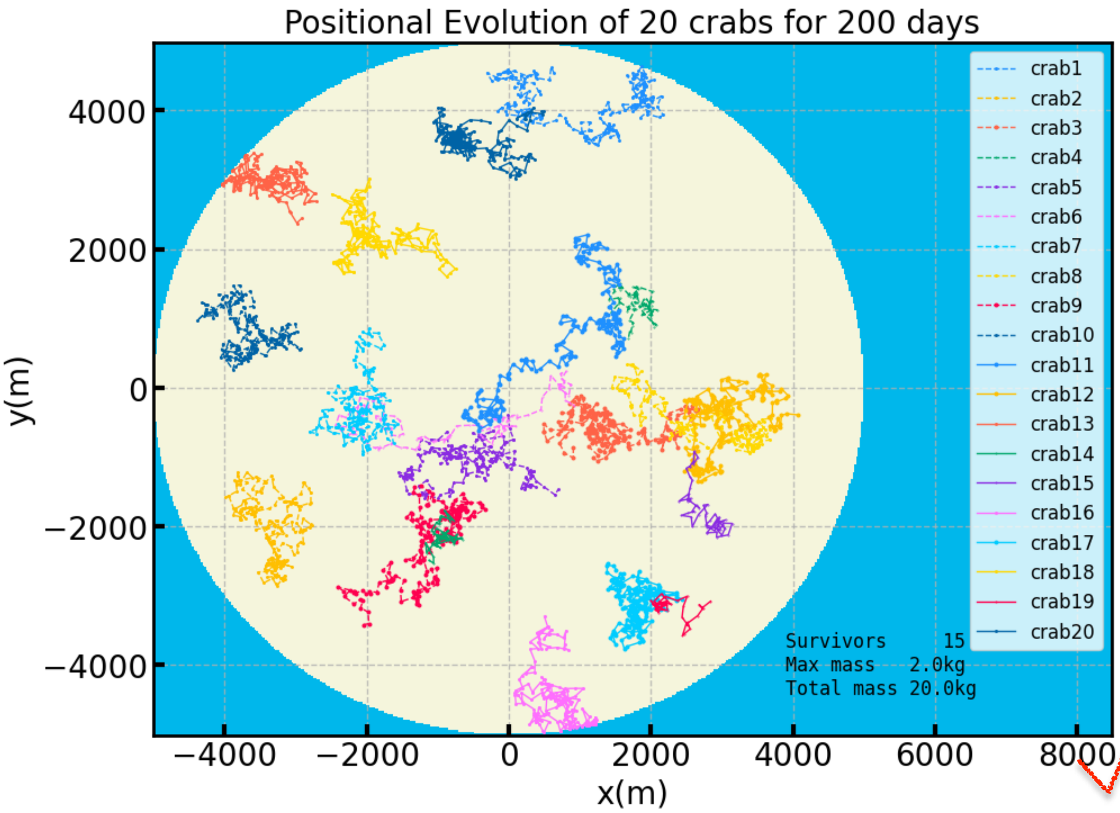


1/1
Good job!

Figure 12: The distribution of the distance of each crab reaches the seashore. The width of each bin is 2km where the total range of the histogram spans from 0 to 200km.

20 Crabs on the island

There are 20 crabs introduced on the island now. Each of these crabs can eat one another to either gain an opponent's mass or expire. I made an observation on the 200th day. Their traces are shown in Fig13. I surveyed the surviving crabs' population and the mass of the most massive crab(s). The 'cancrinegraphic' of the survived crabs in Fig14 and Fig15 shows that the most probable number of the survivors is 16. While the most probable mass of the most massive crab seems to be 3kg. The chances of having a crab that committed more than three predations are not quite notable.



Excellent plot!

Figure 13: The history of a set of 20 crabs on the island. The more massive crabs are marked with larger markers but there were no single outstandingly massive cab on the 20th day.

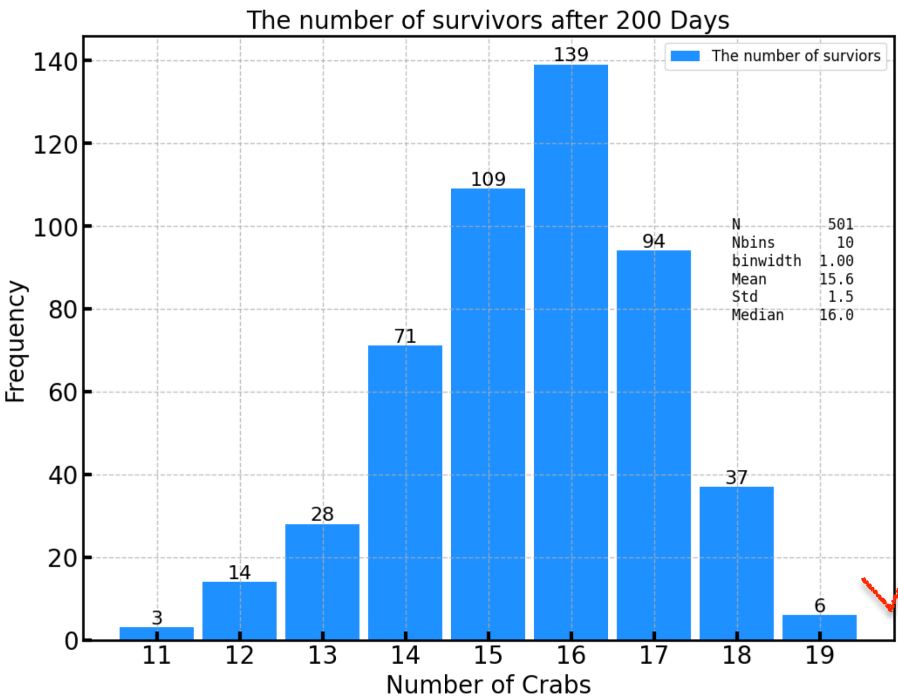
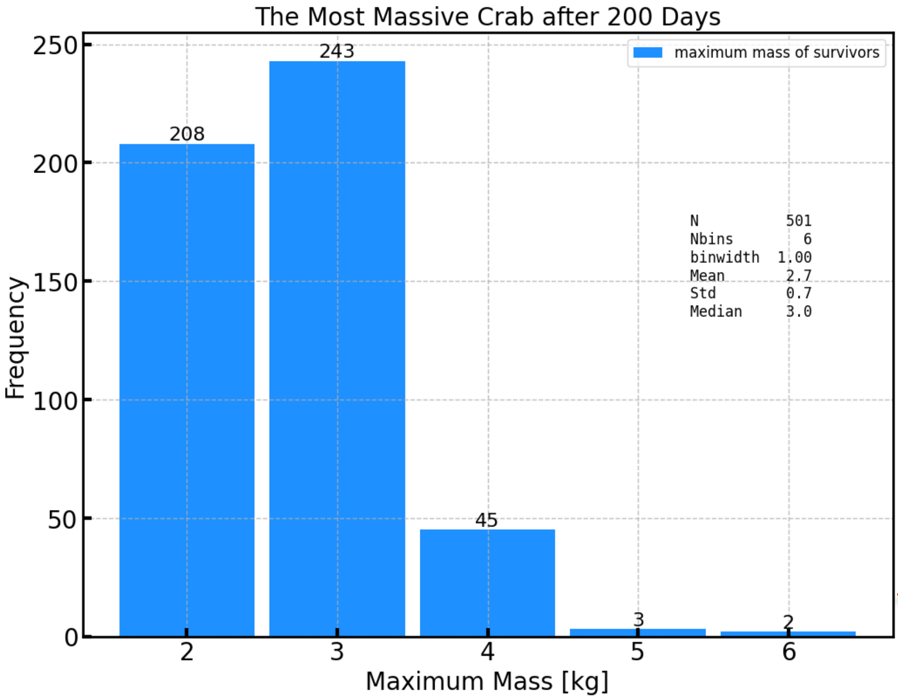


Figure 14: The histogram visualizes the results of 501 experiments tracking crab survivor numbers after 200 days. The majority of experiments resulted in 16 survivors, which is also the median value of the dataset.



1.5/1.5
Excellent!

Figure 15: The histogram depicts the distribution of the maximum mass of crab survivors after 200 days. It shows how often each mass category occurred among the 501 recorded experiments.

Days until having 10 survivors

An ensemble of experiments is prepared progressing until 10 survivors are remaining. The distribution does not have a symmetric feature to its central region. So calculation of the confidence interval requires additional consideration in addition to the mean or median.

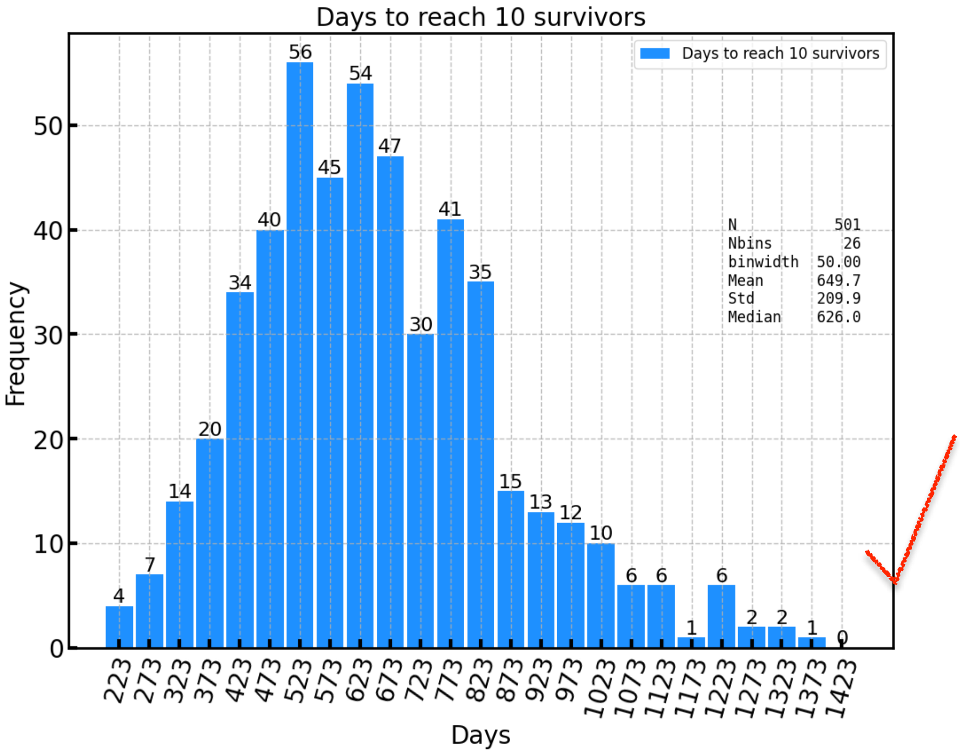
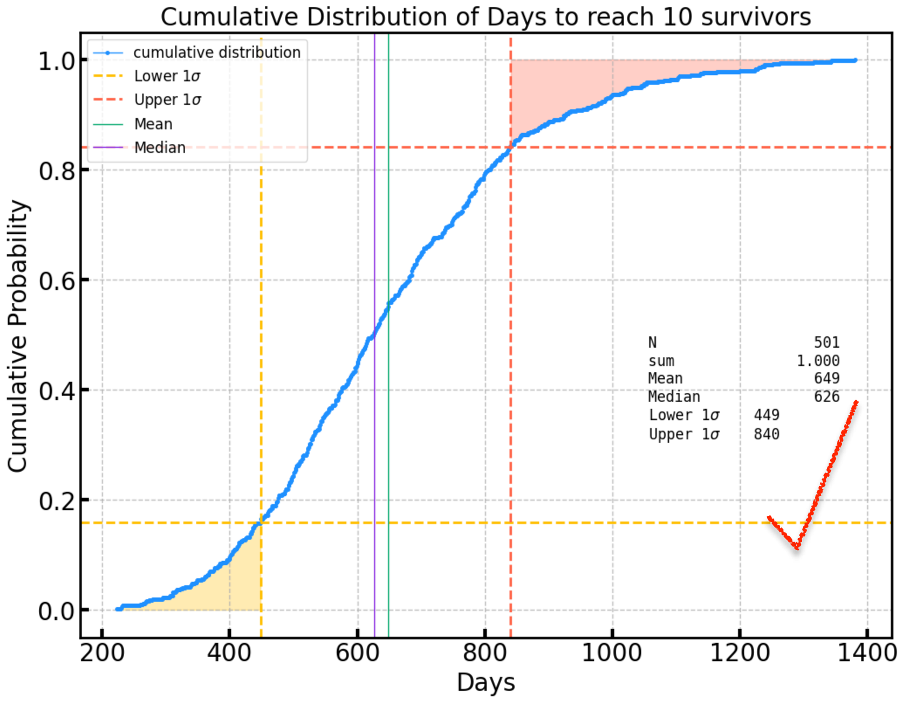


Figure 16: The histogram shows the distribution of the days required to reach 10 surviving crabs. The data spans a range from fewer than 250 days to over 1400 days. The peak of the distribution appears to occur in the 600–650 day range, where the highest frequency of 56 is observed.

1σ confidence interval

The x value of experiments is the number of days which is inherently a discrete value, allowing for the construction of a histogram with a bin width of one. By incorporating all experimental results, and upon normalising the histogram, it can be treated as a probability distribution, which in turn can be approximated as a PDF as long as I have done a sufficient number of experiments. Using that, I get the cumulative probability distribution where I can directly look up the 15.865% quantile of each tail.



1/1
Perfect

Figure 17: The figure depicts the cumulative distribution of the number of days taken to reach 10 survivors in a series of 501 experiments. Due to its asymmetry, the lower 1σ quantile and the upper 1σ quantile are different.

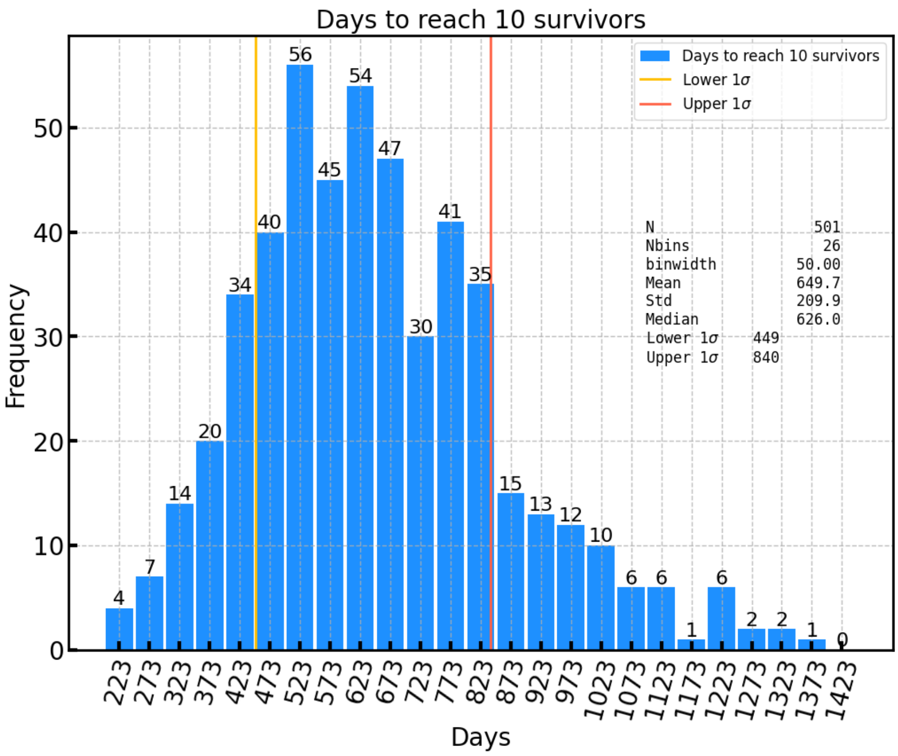


Figure 18: The histogram shows the distribution of the days required to reach 10 surviving crabs. Now lower and upper 1σ quantile is also marked. Due to the asymmetry of its distribution. The 1σ quantile for the lower and upper tails are asymmetric as well.

Problem 3 Total: 0.5+1+1.5+1 = 4